Introduction

This book aims to provide a comprehensive, philosophically grounded, defense of the use of social welfare functions as a framework for evaluating governmental policies and other large-scale choices.

The "social welfare function" (SWF) is a concept that originates in theoretical welfare economics. It is employed as a policy-analysis methodology in a number of economic literatures, such as "optimal tax" scholarship, growth theory, and environmental economics. But other methodologies –in particular, cost-benefit analysis (CBA) – are currently dominant. While CBA is defensible as a rough proxy for overall well-being,[1] it is insensitive to the distribution of well-being. By contrast, the SWF approach can incorporate distributive considerations into policy analysis in a systematic fashion.

Although I see SWFs as a practical policy-evaluation tool, the tenor of this book is theoretical. Just as the now-massive body of CBA scholarship is grounded in a theoretical literature regarding CBA, so, too, the proper design of the SWF framework raises many questions of normative theory – questions that this book will engage. In doing so, I draw upon welfare economics, social choice theory, and related formal literatures (such as utility theory and decision theory), and upon philosophical scholarship concerning a variety of topics, in particular well-being, equality, and personal identity.

Chapter 1 sets the stage. I see the SWF framework as a *moral* choice-evaluation framework. "Moral" reasoning is the species of normative reasoning characterized by a concern for human interests; by impartiality between different persons; and by a willingness to transcend and criticize existing social norms. SWFs provide a systematic tool for *morally* evaluating governmental policies and other large-scale choices. Chapter 1 explores the difference between moral evaluation and other kinds of normative evaluation, and briefly reviews questions of metaethics and normative epistemology that no work of normative theory can ignore. It also sets forth the basic argumentative strategy of this book: to take as given that a moral choice-evaluation framework should be *person-centered, consequentialist* and *welfarist* (for short, "welfarist") and to argue that the SWF approach is the most attractive framework of this sort.

In other words, this book works *within* welfarism, rather than engaging ongoing debates between welfarists and non-welfarists. Chapter 1 explains why this is a plausible strategy. However, it also takes some pains to explain why non-welfarists, too, should find the book of interest.

Chapter 1 concludes by offering a formal, generic, architecture for welfarism. The generic welfarist architecture derives a ranking of choices from a ranking of outcomes. The ranking of outcomes, in turn, depends upon individual well-being. The connection between the

[1] See Matthew D. Adler and Eric Posner, *New Foundations of Cost-Benefit Analysis* (2006).

ranking of outcomes and individual well-being is formalized via the concept of a "life-history": a pairing of a person and an outcome. Life-history (x; i) means being individual i in outcome x. A welfarist choice-evaluation framework includes an account of well-being, which at a minimum makes *intrapersonal* comparisons, ranking life-histories belonging to the same person. The Pareto principles constrain the ranking of outcomes – requiring it to be consistent with the intrapersonal ranking of life-histories in certain, basic, ways.

The SWF approach is one *specification* of this generic welfarist architecture; CBA is a competing specification.

Chapter 2 introduces the SWF framework. This approach has the distinctive feature of making *interpersonal* comparisons between life-histories – not just intrapersonal comparisons. Further, it employs a utility function (or set of such functions) to map each outcome onto a "vector" or list of numbers, representing the well-being of each individual in the population in that outcome. Outcome x is mapped by utility function $u(.)$ onto $(u_1(x), u_2(x), \ldots, u_N(x))$, where $u_i(x)$ is a numerical measure of the well-being of individual i in outcome x. A SWF, in turn, is a mathematical rule for ranking outcomes as a function of their corresponding utility vectors. One simple possibility is to add up utilities: this is the utilitarian SWF. Another possibility is to employ an outcome-ranking rule which is sensitive to the distribution of utilities. There turn out to be a multiplicity of such distribution-sensitive SWFs.

Chapter 2 explains these ideas, and also reviews the intellectual history of the SWF approach (which originates in work by Abram Bergson and Paul Samuelson some 70 years ago, and, as mentioned, is well-accepted within certain subfields of economics). The bulk of the chapter, however, focuses on criticizing the competing policy-analytic frameworks that are currently dominant. These competitors include not only CBA, but also inequality metrics, such as the well-known Gini coefficient; various other types of metrics for quantifying inequity, such as poverty metrics, "social gradient" metrics, and tax incidence metrics; and cost-effectiveness analysis (CEA). Each of these approaches is widely employed in academic work, and CBA also now has a firm legal status in governmental practice. However, each of these approaches is problematic – at least from the perspective of welfarism.[2] As Chapter 2 will show, these approaches may be vulnerable to violations of the Pareto principles, or may fail to rank outcomes in a well-behaved manner (for example, by ranking outcome x over y but y over x, or x over y and y over z but not x over z). And even if non-SWF methodologies *are* structured so as to yield a well-behaved, Pareto-respecting ranking of outcomes, they turn out to be problematic in other ways.

The analysis in Chapter 2 is meant to *motivate* the defense and elaboration of the SWF approach which occurs in subsequent chapters. Chapters 3, 4, and 5 address the central theoretical questions that must be confronted by any proponent of this approach. Chapter 3

[2] Alternatively, certain ways of employing currently dominant frameworks turn out to be variations on the SWF approach. This is true, in particular, of the use of CBA with so-called "distributive weights." See Chapter 2.

focuses on well-being. One philosophically contested issue concerns the choice between preferentialist, hedonic, and objective-good accounts of human welfare. Insofar as utility numbers are meant to quantify individual well-being in outcomes, what exactly should these numbers be measuring? A cross-cutting issue concerns interpersonal comparability. How are we to make sense of the statement that life-history (x; i) is better for well-being than life-history (y; j): that individual i in outcome x is better off than individual j in outcome y? Why believe that this statement is meaningful? What are the criteria for ranking life-histories involving different persons? Economists outside the SWF tradition are usually skeptical about the possibility of interpersonal comparisons. Many SWFs also make interpersonal comparisons of well-being *differences*, saying that the difference in well-being between life-history (x; i) and (y; j) is greater than the difference in well-being between life-history (z; k) and (w; l). But what are the criteria that would enable us to make sense of *these* sorts of comparisons?

Chapter 3 tackles these problems, proposing to analyze well-being in terms of fully-informed, fully rational, convergent extended preferences. While an ordinary preference is simply a ranking of outcomes and choices, an *extended preference* is a ranking of life-histories. To say that individual k has an extended preference for (x; i) over (y; j) means that k prefers the life-history of i in x to the life-history of j in y. The idea of an extended preference originates with John Harsanyi. More specifically, Harsanyi proposes that an interpersonally comparable metric of individual well-being be constructed by appealing to individuals' extended preferences over life-history *lotteries* – on the premise that these extended lottery preferences comply with expected utility theory. Chapter 3 will develop Harsanyi's fruitful ideas. To be sure, many challenges arise in doing so; and the account of well-being presented in Chapter 3, in a number of important respects, diverges from Harsanyi's views. In particular, my definition of extended preferences builds in a self-interest component, designed to screen out preferences for features of outcomes that have no impact on well-being; and I allow for heterogeneity in extended preferences.

The thrust of Chapter 3 is to defend the following approach for making intra- and interpersonal comparisons, and for measuring well-being via utility numbers. There is a set **U** of utility functions, pooling the fully informed, fully rational, extended preferences of everyone in the population. Life history (x; i) is at least as good for well-being as life-history (y; j) just in case $u(x; i) \geq u(y; j)$ for all $u(.)$ in **U**. A similar rule is proposed for well-being differences.[3]

Chapter 4 turns to the question of specifying the SWF. An SWF is some rule for *using* the well-being information captured in the set **U** of utility functions in order to rank outcomes. Chapter 4 argues that the most attractive such rule is a *prioritarian* SWF (more precisely, a "*continuous* prioritarian" SWF). In defending this view, Chapter 4 draws heavily on the contemporary philosophical literature concerning equality. One major theme in this literature is

[3] The well-being difference between life-history (x; i) and (y; j) is at least as great as the well-being difference between life-history (z; k) and life-history (w; l) iff, for all $u(.)$ in **U**, $u(x; i) - u(y; j) \geq u(z; k) - u(w; l)$.

the debate between those who hold a "prioritarian" conception of fair distribution, and those who reject this view. "Prioritarians" argue that well-being changes affecting worse-off individuals have greater moral significance. In other words, well-being has declining marginal moral weight. It is this proposition, and not the intrinsic value of equality, that provides the best justification for a non-utilitarian moral view – or so "prioritarians" claim. Prioritarianism corresponds to an SWF which satisfies two key axioms, explained in Chapter 4: the "Pigou-Dalton" axiom, and an axiom of separability across persons. If we add a continuity requirement, the upshot is a SWF which sums up individual utilities that have been "transformed" by a transformation function, rather than simply summing utilities in utilitarian fashion.

Formally, a continuous prioritarian SWF says: outcome *x* is morally at least as good as outcome *y* iff, for all *u*(.) belonging to **U**, $\sum_{i=1}^{N} g(u_i(x)) \geq \sum_{i=1}^{N} g(u_i(y))$, where the *g*(.) function is strictly increasing and concave (which is what ensures that this SWF both satisfies the Pareto principles *and* gives greater moral weight to well-being changes affecting worse-off individuals). Chapter 4 argues that this SWF represents the most attractive specification of welfarism.[4] A central claim in Chapter 4, and indeed throughout the book, is that welfarism and a concern for *fairness* are fully compatible. A moral view is sensitive to fairness insofar as it "respects the separateness of persons" – insofar as it sees each person as having a separate moral claim to have her interests and concerns respected. Integrating a concern for fairness into welfarism means, first, that fairness structures the ranking of *outcomes*; and, second, that the "currency" for each individual's moral claim is her well-being. These ideas, in turn, lead most directly to the prioritarian SWF. Whether the prioritarian SWF should, in addition, satisfy the continuity requirement – and I believe it should – implicates questions regarding tradeoffs that are also reviewed in Chapter 4.

Chapter 5 addresses the temporal dimension. The approach defended in Chapters 3 and 4 makes the ranking of outcomes depend upon utility numbers representing individuals' lifetime well-being. A whole-lifetime view is, indeed, adopted by the theoretical literature on SWFs; by most extant scholarship that uses SWFs to evaluate governmental policies; and by the philosophical literature on equality, which generally argues that moral norms concerning fair distribution are properly focused on the distribution of lifetime well-being. But is the whole-lifetime approach really defensible? Why not represent an outcome as a list of "sublifetime" utilities, each representing the well-being of some individual during some portion of her life (for example, her annual or momentary well-being), and then apply a continuous prioritarian SWF to these sublifetime utilities? Chapter 5 will describe and seek to respond to two arguments that challenge whole-lifetime prioritarianism, and that seem to cut in favor of "sublifetime"

[4] More precisely, Chapter 4 argues for the "Atkinsonian" SWF, which is a particular type of continuous prioritarian SWF – and one that, in fact, is fairly widely used within existing SWF scholarship.

prioritarianism or some other approach.[5] One argument, tendered by Derek Parfit, suggests that a proper understanding of personal identity undercuts a concern for the distribution of lifetime well-being. A different argument, advanced by Dennis McKerlie and other philosophers, suggests that our intuitions about equality – in particular, intuitions about the moral significance of short-term hardship and suffering – are inconsistent with a whole-lifetime view.

Chapter 6 turns to the problem of implementation. While Chapter 3 undertook the philosophical labor required to defend a particular theory of well-being and well-being measurement, the question remains: how shall we actually estimate the utility functions which the SWF approach requires as its inputs? How shall we actually construct a set **U**? Chapter 6 addresses this question at length. It begins by addressing a question left open in Chapter 3. The outcomes which are ranked by a choice-evaluation framework (be it the SWF framework or a competing framework, such as CBA) are *simplified* descriptions of reality. Simplification is necessary for the framework to be cognitively tractable. (If an outcome were a fully precise specification of a possible reality, i.e., a complete "possible world," a *human* decisionmaker would be unable to use the framework.) But what does it mean for individuals to have extended preferences regarding life-histories involving simplified outcomes – outcomes that are missing some characteristics? For example, much SWF scholarship in the "optimal tax" tradition employs outcomes that describe each individual's consumption and leisure, but fail to describe other individual attributes (health, happiness, social life, etc.). How should individual *k* think about her preference regarding (*x*; *i*) and (*y*; *j*), where she is told only that individual *i* consumes a certain amount and has a certain amount of leisure time in outcome *x*, and that individual *j* consumes a certain amount and has a certain amount of leisure time in outcome *y*?

Chapter 6 proposes an answer to this vital question, regarding the valuation of simplified outcomes. [6] With that answer in hand, it discusses how we can use information about an individual's ordinary preferences in order to make inferences about her extended preferences. And it reviews, in detail, the wealth of existing data concerning individuals' ordinary preferences that enable a policy analyst to construct a set **U**: data regarding individuals' preferences for consumption lotteries; evidence concerning intertemporal substitution and the value of statistical life; "ordinal" preference data supplied by economic research concerning labor supply and consumer demand; so-called "QALY" surveys, which reveal how individuals rank health states and lotteries over health states; and happiness surveys. This chapter also proposes novel survey formats.

[5] A third approach would be "attribute based," whereby an SWF is applied directly to individual attributes, rather than to lifetime or sublifetime utilities representing individuals' lifetime or sublifetime well-being.

[6] The answer, in short, is that the enterprise of eliciting individuals' preferences with respect to simplified life-histories rests upon an *invariance* premise: that such preferences are more or less invariant to the particular level of the missing characteristics. If the invariance premise is untrue, a fuller description of outcomes is warranted – although considerations of cognitive tractability will weigh against describing outcomes with great specificity.

Chapters 3 through 6 all focus on the ranking of outcomes. Is the well-being of a given individual in a given outcome determined by her preference-satisfaction, her mental states, or her realization of objective goods? How should her well-being be measured by utility functions? What sort of data enables us to estimate these functions? What is the appropriate SWF for ranking outcomes in light of individual utilities? Is it a utilitarian SWF, a prioritarian SWF, or some other form?

Chapter 7 turns from these questions, to the problem of generating a ranking of choices from the ranking of outcomes. A choice-evaluation framework should function to provide guidance to a decisionmaker. In particular, the SWF framework – as I conceptualize it – is a systematic methodology that should yield guidance to governmental policymakers or others confronted with large-scale choices.[7] But a human decisionmaker operates under conditions of uncertainty. She is not sure which particular outcome would result from any given choice which is available to her. How to implement a continuous prioritarian SWF under conditions of uncertainty raises thorny problems. It turns out that no methodology for doing so can simultaneously respect, on the one hand, certain axioms which seem to capture the essence of consequentialism; and, on the other hand, the ex ante versions of the Pareto and Pigou-Dalton principles.

Expected utility theory, if refined along certain lines, provides an attractive generic structure for choice under uncertainty.[8] Chapter 7 argues that a continuous prioritarian SWF should be merged with a (refined version of) expected utility theory so as to generate a ranking of choices – notwithstanding violations of the ex ante Pareto and Pigou-Dalton principles. While the SWF framework defended here satisfies the Pareto and Pigou-Dalton principles in terms of the ranking of outcomes, the ex ante versions of these principles constitute an *additional* requirement which, on balance, should be rejected. The dilemmas that arise in specifying norms of fair distribution under conditions of uncertainty have been discussed by philosophers and social choice theorists; Chapter 7 builds upon this scholarship.

Chapter 8 reviews three important problems that are connected to those addressed in this book. It describes the problems, and in a very limited way outlines tentative responses, but does not attempt anything like a full treatment.

One problem concerns future generations. My analysis throughout the book assumes a fixed and finite population. The same *N* individuals exist in each of the possible outcomes of the policy choice at hand. Scholarship on future generations relaxes this assumption, by

[7] Policy-evaluation frameworks such as the SWF approach, CBA, or the other frameworks reviewed in Chapter 2 are appropriate for governmental policies or other large-scale choices, but not for smaller choices where the expected benefits of using a systematic framework are too small to justify the decision costs of doing so. Identifying the boundary between small and large choices is very difficult. See Chapter 1.

[8] To be clear, expected utility theory surfaces at two different junctures in this book: in Chapter 3, regarding the measurement of well-being; and in Chapter 7, regarding the moral ranking of choices under conditions of uncertainty.

allowing for the possibility that choices might affect the size or identity of the population, or for an infinite future and thus infinite population. How to structure policy choice under such conditions raises new and difficult questions: the so-called "repugnant conclusion"; "non-identity" problems; and the incompatibility between the Pareto principles and an axiom of impartiality in the case of an infinite future.

A second problem concerns the optimal design of legal institutions. To say that the SWF approach is an attractive framework for morally evaluating governmental policies and other large-scale choices is not, necessarily, to say that it is optimal to structure legal institutions so that policymakers are legally instructed to employ this framework. Policy-analysis tools may be distorted by political forces. (In particular, research examining the effects of laws requiring regulatory agencies to employ CBA has reached mixed conclusions concerning whether such laws have actually produced more efficient regulations.) A cross-cutting idea is that it may be optimal to "channel" distribution through the tax system, and thus to instruct non-tax bodies to evaluate their decisions using CBA rather than using some SWF which is sensitive to distributive considerations.

A third problem concerns individual responsibility. A key deficit of welfarism is that it fails to differentiate between bad luck and irresponsibility -- between a case in which someone is badly off through no fault of her own, and a case in which someone is (wholly or partly) responsible for her well-being shortfall. Over the last several decades, the philosophical literature on equality has intensively investigated problems of responsibility; and a growing body of work in welfare economics and social choice theory is now also engaging such problems. Chapter 8 briefly reviews these literatures, and in a preliminary way suggests how a concern for responsibility might be fused with the SWF framework.

This book is, evidently, interdisciplinary. It is aimed at welfare economists who are receptive to philosophical argumentation; at philosophers who are receptive to the mathematical tools of welfare economics; and to law and policy scholars who find value in both fields. It builds upon, and draws inspiration from, the tradition of scholarly work at the intersection of philosophy and economics, exemplified by journals such as *Economics and Philosophy* or *Social Choice and Welfare*. The methodology of welfare economics is axiomatic and deductive. The focus is on clarifying the logical implications of various axioms for ranking outcomes and choices which we might be inclined to endorse. The methodology of moral philosophy is coherentist. Given a plurality of logically possible approaches to ranking outcomes and choices, which ones are most attractive in the "reflective equilibrium" sense? Which approaches fit best with our intuitive judgments about concrete cases and with general normative principles, regarding well-being, equality, and so forth?

It would be arrogant and wrongheaded to suggest that normative understanding can only be advanced by interdisciplinary work. Clearly, that is not true; there are large epistemic gains to be had from specialization. However, it seems to this author equally wrongheaded to insist that

specialization is the only viable path. This book is animated by the belief that scholars can make real progress in specifying normative tools and frameworks by marrying the methodologies of economics and philosophy. I'll leave it to the reader to judge whether they are, in fact, fruitfully married here.

Chapter 1. Preliminaries: Morality, Consequentialism, Welfarism

[The first part of chapter 1, which addresses metaethical and epistemological issues, and explains why I work within welfarism, is not included. What follows is the section of the chapter that provides a more formal characterization of welfarism.]

Welfarism: A Formal Characterization

Prior sections have reviewed the case for why a moral decision procedure should be "welfarist" – that is, exclusively person-centered, consequentialist, and oriented around human well-being rather than other possible currencies for ranking outcomes. Here, I provide a more rigorous characterization of the structure of a welfarist decision procedure. This formal structure will be the basis for discussions throughout the rest of the book.

There is some decisionmaker, who at some moment in time t^* is trying to decide what she morally ought to do at t^*. A moral decision procedure or evaluation framework consists in methodologies, formulas, principles, etc., that the decisionmaker can actually use in arriving at a conclusion about what she should do.[9]

A welfarist decision procedure, because it is consequentialist, helps the decisionmaker answer this question via the following structure. It tells the decisionmaker to construct a choice set $\mathbf{A} = \{a, b, c \ldots\}$, comprising some of the actions that the decisionmaker might take; it tells the decisionmaker to construct an outcome set $\mathbf{O} = \{x, y, \ldots\}$, comprising some of the possible outcomes of the choices; it furnishes principles for morally ranking the outcomes in **O**; and it furnishes principles for deriving a moral ranking of the choices in **A** from the ranking of the outcome set **O**.

More precisely, the decision procedure tells the decisionmaker to reason in this explicitly consequentialist fashion for sufficiently "large scale" choices. I will not discuss what the form of normative guidance for small-scale choices is, nor do I have any insight into the threshold procedure that the decisionmaker should use in deciding whether the choice at hand is "large-scale."

An important debate in decision theory concerns whether the elements of a choice set **A** should be thought of as immediate actions (physical actions or speech acts that the decisionmaker can perform at t^*) or intertemporal plans (plans which the decisionmaker can adopt at t^*, whereby she formulates an intention to perform various physical actions or speech acts at some points in time after t^*, typically contingent on what occurs after t^*). In the case of a

[9] In addition, a researcher or adviser can presumably use it to evaluate what the decisonmaker morally should do; and these third parties, or the decisionmaker herself, can presumably use the procedure at some time before t^* or after t^* in order to decide what the decisionmaker morally should do or should have done at t^*. A welfarist choice evaluation procedure will have the very same structure in these cases of third-party or noncontemporaneous assessment as in the paradigm case of first party, contemporaneous assessment, although the particular probabilities assigned to outcomes (in moving from the outcome to the action ranking) will be different.

policymaker, an immediate action might be making a legal utterance (enacting a regulation or statute) at *t**, while a temporally extended plan might be formulating an intention to make a series of legal utterances over time. This question is discussed in Chapter 7. The question is, obviously, a very important one – but it can be detached from questions concerning the construction and ranking of the outcome set **O**, which will be the focus of Chapters 2 through 6. The analysis and argumentation in those chapters is fully consistent both with the view that the objects of choice are immediate actions, and with the view that they are intertemporal plans.

The decisionmaker is a human decisionmaker, and thus her cognitive abilities are limited. She is "boundedly rational." An attractive moral decision procedure must be one that *she* can use. This means, to begin, that the choice set **A** cannot consist of every possible immediate action that the decisionmaker can perform at *t**, or every possible plan she can formulate at *t**. It is not within the cognitive capacities of a human decisionmaker (even aided by computers) to think about the totality of her possible choices. So **A** must consist in some subset of the immediate actions/plans available to the decisionmaker at *t**. She consciously considers some alternatives, and just ignores others. How to identify that subset is a huge problem in decision theory – not just for welfare consequentialists, but for anyone who purports to give normative guidance – and little intellectual progress has been made in solving it. I will not try to tackle the problem in this book, but will rather assume that the decisionmaker has somehow constructed a cognitively tractable set of alternatives **A**, and is now trying to rank its elements as morally better or worse.

The bounded rationality of the decisionmaker means, too, that the outcomes in **O** cannot be complete "possible worlds." A possible world is a full description of a possible history of the universe, from start to finish. It is beyond the cognitive abilities of a human decisionmaker (even aided by computers) to contemplate or mentally process a single possible world, let alone a set of them. Thus each outcome in **O** will be a limited, incomplete description of a possible world.[10]

[10]It is sometimes argued that the "outcomes" which figure in consequentialist analysis should not be whole possible worlds (or simplified descriptions thereof), but *future* outcomes (descriptions of possible future histories of the world), or only those states of affairs that would be *caused* by the choice at hand. As for the first possibility: to the extent that the moral value of future outcomes is not separable from what has occurred in the past, why shouldn't the outcomes that figure in moral decisionmaking potentially describe the past as well as the future? (In point of fact, because a prioritarian SWF cares about the distribution of individuals' lifetime well-being, the *moral* ranking of a given set of possible future outcomes will generally *not* be separable from what has occurred in the past – even if each individual's lifetime utility function, measuring his lifetime well-being, is temporally separable.) As for the second possibility: The moral value of those states of affairs that would be caused by the choice at hand may not be separable from states of affairs which obtain independently. Thus ignoring the latter may skew choice. The best rationale for "causal conseqentialism" has to do with a worry concerning choice under uncertainty – namely, that the increase in probability of some outcome *x* which is associated with performing action *a* rather than *b* should reflect the fact that *a* is likelier to cause *x*, and not merely the fact that *a*'s performance is stronger evidence that *x* will obtain. But this important worry is handled by the apparatus of causal decision theory (for example, via the Savage set up, which assigns probabilities to outcomes by assigning them to "states" which are causally independent of the choices at hand). See Chapter 7. If this apparatus is employed, outcomes should include everything that matters to

More specifically, each outcome will have a single period or multiple periods. A period is some length of time: a day, year, decade, etc. The generic characterization of each period in a given outcome is $(\mathbf{a}_1^t, \mathbf{a}_2^t, ..., \mathbf{a}_N^t, \mathbf{a}_{imp}^t)$, where $\mathbf{a}_i^t$ is a vector of attributes of individual *i* during period *t*, and $\mathbf{a}_{imp}^t$ are background facts about the world, such as the price vector, environmental quality, causal laws, or any other facts that are not attributes of individual persons. The individual attributes included in $\mathbf{a}_i^t$ will be a subset of the totality of types of attributes that characterize persons in a full possible world; similarly, background facts will be a incomplete description of the background facts that obtain in any possible world.

For example, an outcome might simply describe each individual's period-specific consumption of *M* marketed goods, so that $\mathbf{a}_i^t$ takes the form $(q_i^1, q_i^2, ..., q_i^M)$, where q_i^m is the quantity of good *m* that individual *i* consumes during period *t*. Or it might simply describe the total money value of each individual's consumption during the period, so that $\mathbf{a}_i^t$ takes the very simple form c_i. In these cases, all of an individual's non-consumption attributes are left undescribed. A different possibility is that an outcome describes an individual's total consumption, leisure, and level of some public good during the period, so that $\mathbf{a}_i^t$ takes the form (c_i, l_i z_i). Alternatively, it might characterize an individual's health state (alone, or in conjunction with other attributes) or his level of happiness.

That the elements of **O** consist of partially described possible worlds should be unsurprising to anyone familiar with policy analysis or, more generally, economic modeling. The "states of the world" or outcomes that are ranked using CBA or other standard policy-analytic frameworks, or that figure in economic models, are always the kinds of limited descriptions of reality discussed in the previous paragraph. (The types of individual attributes included in these simplified outcomes are very often some subset of the types mentioned in the previous paragraph: an individual's consumption, leisure, health, happiness, and enjoyment of different public goods.) The simplified character of the elements of **O**, and what it means to apply a social welfare function to such an outcome set, is discussed at great length in Chapter 6.

The decision procedure derives a ranking of the choice set **A** from a ranking of the outcome set **O**. More precisely, this ranking is a quasiordering, possibly incomplete. Throughout the book, when I use terms such as "ranking," "ordering," and so forth, I mean a possibly incomplete quasiordering, unless I specify otherwise.

What is a "quasiordering"? This is a very general idea, which can be used in talking about any set of objects **S** = {*q*, *r*, *s*, ….}. A quasiordering is a *reflexive* and *transitive* binary relation on the set. It is typically denoted by the symbol "$\succcurlyeq$" or the phrase "at least as good as."

---

the decisionmaker (modulo the need to simplify outcomes so as to make them cognitively tractable), not merely causal consequences.

The relation is *reflexive*: Each object is at least as good as itself. The relation is *transitive*: If q is at least as good as r, and r is at least as good as s, then q is at least as good as s.

A quasiordering can, but need not, be complete. A complete quasiordering has the following property: For every possible pair of items in **S**, either the first is at least as good as the second, or the second is at least as good as the first, or both. An incomplete quasiordering lacks this completeness property. That means that there is at least one pair of items, call them s and t, such that s is not at least as good as t, nor t at least as good as s. In this case, it is typical to say that s and t are *incomparable* or *noncomparable*.

Given some set **S** and some reflexive and transitive binary relation on that set (the "at least as good as" relation, $\succcurlyeq$), we can define two other relations: a *transitive, irreflexive,* and *asymmetric* relation denoted "$\succ$" or "better than"; and a *transitive, symmetric,* and *reflexive* relation denoted "=" or "equally good as." Given two items in **S**, the first is better than the second iff (1) the first is at least as good as the second and (2) the second is not at least as good as the first. Given two items in **S**, the first is equally good as the second iff (1) the first is at least as good as the second, and (2) the second is at least as good as the first. It is easy to see that the binary relations of "better than" and "equally good as," defined in this way from the basic relation of "at least as good as," are indeed transitive, irreflexive and asymmetric (in the case of "better than") and transitive, reflexive and symmetric (in the case of "equally good as").

Strictly speaking, a quasiordering is any reflexive, transitive binary relation $\succcurlyeq$ on a set **S**, and the relations $\succ$ and = are just the relations generated from $\succcurlyeq$ in the manner just described. But it is extremely natural to use the natural-language term "at least as good as" to refer to $\succcurlyeq$, and the terms "better than" and "equally good as" to refer to $\succ$ and =, respectively. It is a platitude about the ordinary concept of "better than" that it is transitive, irreflexive, and asymmetric; it is a platitude about the ordinary concept of "equally good as" that it is transitive, reflexive, and symmetric. Intuitively (at least if one is a consequentialist) it seems that our decisions should reflective the comparative "goodness" of outcomes and actions, in the ordinary sense of "good," and it is a natural thought to formalize this concept via the relations $\succcurlyeq$, $\succ$ and =.

As I've said, a quasiordering is a very general notion. In the case of a welfarist decision procedure, as I conceptualize it, there are three kinds of sets that are associated with quasiorderings: the outcome set **O**, the action set **A**, and a set **H** of life-histories (an idea I'll get to in a moment.)

The decision procedure will provide tools for arriving at a quasiordering of the outcome set **O**. In other words, it will provide tools for associating a reflexive, transitive, possibly incomplete binary relation with **O**, the binary relation "morally at least as good as," which I will abbreviate $\succcurlyeq^{M}$. Each outcome will be morally at least as good as itself. If x is morally at least as good as y, and y is morally at least as good as z, then x is morally at least as good as z. The relation may, but need not, be complete. (If incomplete, there will be at least one pair of

outcomes, *x* and *y*, such that *x* is not morally at least as good as *y* nor y morally at least as good as *x*.) Once the decision maker has employed the furnished tools to arrive at the $\succcurlyeq^{M}$ relation, she can immediately construct two other relations, "morally better than" and "morally equally good as," which we can denote as $\succ^{M}$ and $=^{M}$, respectively. [11]

The decision procedure will also provide tools for using the quasiordering of the outcome set **O**, to arrive at a quasiordering of the action set **A**. This will be a reflexive, transitive, possibly incomplete binary relation of the *actions* in that choice set, ranking some as "morally at least as good as" others. Denote this quasiordering as "$\succcurlyeq^{MA}$". With the moral quasiordering of the action set in hand, at least if that set is finite, it is straightforward for the decisionmaker to determine what he morally should do: He should morally perform some "undominated" action (one which is not morally worse than any other), and is permitted to perform any one. [12]

What about the derivation of the moral quasiordering of the action set, $\succcurlyeq^{MA}$, from the moral ordering of the outcome set, $\succcurlyeq^{M}$? In the case of choice under certainty – where the decisionmaker knows which outcome any given action would produce – this derivation is trivial. The decision procedure simply says this: if action *a* yields outcome *x* and action *b* yields outcome *y*, then $a \succcurlyeq^{MA} b$ just in case $x \succcurlyeq^{M} y$. In the more realistic case of choice under uncertainty, the process of deriving the moral quasiordering of the action set from the moral quasiordering of the outcome set is much more complicated and contestable – involving the assignment of probabilities to outcomes, and different possible ways to integrate this probability information with $\succcurlyeq^{M}$ to derive $\succcurlyeq^{MA}$. This topic is the focus of Chapter 7.

In particular, Chapter 7 will discuss the use of expected utility (EU) theory -- the best developed tool for moving from a ranking of outcomes to a ranking of choices, under conditions of uncertainty. The applicability of EU theory to the particular case of *moral* choice is controversial. Even if EU theory is applicable in this case (as I will indeed argue), it may need to be refined in various ways.

Another complication, discussed in Chapters 6 and 7, is that the decision procedure may furnish the decisionmaker with principles for constructing a quasiordering of the outcome set **O**, but the decisionmaker may be uncertain how to implement those principles. For example, the decision procedure may tell the decisionmaker that whether *x* is at least as good as *y* depends on individuals' fully informed, fully rational preferences – but the decisionmaker may not know what individuals' fully informed, fully rational preferences are. In this event, the decision procedure – in order to function as a genuine guide to choice – will need to contain methodologies that allow the decisionmaker to integrate the basic principles about how to quasiorder **O**, together with her imperfect information about how to implement those principles,

[11] Again, $x \succ^{M} y$ iff (1) $x \succcurlyeq^{M} y$ and (2) not $y \succcurlyeq^{M} x$. And $x =^{M} y$ iff (1) $x \succcurlyeq^{M} y$ and (2) $y \succcurlyeq^{M} x$.

[12] If the action set is infinite, there may not be any such undominated action – raising complications I will not pursue. These complications are not a result of incompleteness. A complete ranking of an infinite set may lack a single element or subset of elements that are at least as good as every other element in the set.

and her imperfect information about which outcomes a given action yields, to arrive at a quasiordering of **A**.

For much of the book, until these latter chapters, the complications posed by decisionmaker uncertainty about the connection between actions and outcomes, and about the implementation of principles for quasiordering outcomes, will be ignored. The focus, instead, will be on developing attractive principles for quasiordering a given outcome set **O**. Such principles are the absolute core (although certainly not the totality) of a consequentialist decision procedure.

The reader may well raise a number of different objections to a decisionmaking structure that allows for a quasiordering (possibly incomplete) on an outcome set, and a quasiordering (possibly incomplete) on an action set.

First, the reader may object that the ordering of outcomes and actions should be complete. Each outcome should be morally better, or worse, or equally good as every other; similarly, each pair of actions should be thus ranked. Indeed, welfare economists and social choice theorists usually assume completeness in the "social" (what I'm calling "moral") ordering of outcomes and policies; economists typically assume that an individual's preference ordering of outcomes and choices is complete; and the generic structure for rational choice provide by EU theory also assumes completeness. So allowing for incompleteness *is* a deviation from – I view it as a refinement of – EU theory.

However, Amartya Sen has powerfully argued that we should allow for incompleteness in various domains, and other prominent scholars have concurred with Sen. It is especially natural for the welfarist to allow for an incomplete moral ordering of outcomes and actions, because a strong case can be made that well-being is incomplete. Those who deny the possibility of interpersonal welfare comparisons espouse a wide-ranging kind of incompleteness: on this view, it is never possible to compare one person's well-being in some outcome to another person's well-being in some outcome. This is (I will ultimately argue) too extreme; but the insistence that each person's well-being in each outcome is *comparable* with every other person's well-being in each outcome goes too far in the other direction. The most plausible ranking of individual lives in terms of well-being is incomplete, or at least may well be.[13] In turn, therefore, to require completeness of the moral ordering of outcomes– an ordering which, for the welfarist, is a function of the pattern of individual well-being -- seems dogmatic. If we can't always make comparisons of individual well-being, why assume, as welfarists, that we can always make moral comparisons of outcomes?

[13] The account of well-being presented in Chapter 3 derives a well-being ranking of life-histories from a set **U**, which is the union of sets $\mathbf{U}^1,...,\mathbf{U}^N$, where $\mathbf{U}^k$ consists of utility functions that represent individual k's extended preferences over life-histories. One life-history is at least as good as another if its utility is at least as great for all utility functions in **U**. Because indivdiuals' extended preferences over life histories may well be heterogeneous, this approach naturally yields some incompleteness in the well-being ranking of life histories.

Another part of the welfarist case for incompleteness involves the Pareto principle. The Pareto principle, itself, only generates an incomplete quasiordering of outcomes. The project of welfarism, in effect, is to "extend" the Pareto principle (see below); but, once more, to insist that it be "extended" all the way to a complete ordering seems too restrictive.

The strongest case for assuming completeness involves the worry that incompleteness may generate "value pumps." This *is* a serious worry, discussed in Chapter 7; but there may well be other ways to handle "value pumps" than by requiring completeness in the action and outcome orderings; and so we should not assume completeness from the outset. Note, more generally, that the device of the quasiordering does not require incompleteness. It simply *permits* an incomplete ordering. In effect, with regards to the nature of the ordering, the structure provided here is generic; the case of a complete ordering is one specification of that generic structure, and the reader can add that stipulation if she believes it justified. [14] Formally, she does this by stipulating that $\succcurlyeq$ is not only reflexive and transitive, but relates every pair of items in the set **S** at issue.

A second objection comes from the reader who accepts that outcomes and actions may be morally incomparable, but objects to the use of the quasiordering as the formal device for capturing such incomparability. There is a substantial literature regarding quasiorderings, both in mathematical logic, and in economics, where this has become the dominant approach to handling incomparability insofar as that is discussed. Other approaches are not equally well developed.[15] Further, the quasiordering has an elegant "intersectional" structure: if there is a set of complete orderings on **S**, then the relation *R* constructed by saying that *s R t* iff *s* is at least as good as *t* according to each of the complete orderings is a quasiordering; and it turns out that the converse is true, namely that every quasiordering can be represented as the intersection of some set of complete orderings. The "intersectional" structure of quasiorderings also means that the standard mathematical apparatus of functions used by economists to represent complete orderings smoothly generalizes to the quasiordering case: if there is a set of complete orderings on **S**, each represented by a function, then the intersection of these functions generates a quasiordering. Finally, it means that the quasiordering is a natural way to represent the kind of

[14] Admittedly, doing so will require certain moves later in the analysis– specifically, requiring individuals' extended preferences to be homogenous.

[15] Some philosophers have argued that two items might be "on a par" or "roughly equal" rather than incomparable. However, the existence of this fourth relation is controversial, and the formal apparatus for handling it is undeveloped (and indeed may be very close to the intersectional apparatus employed here).

Another possibility, argued by John Broome, is that apparent incomparability is really a form of vagueness. Broome's account, too, has not been widely adopted. One count against it is that vagueness, standardly, is a feature of language, not properties. It might be countered that normative properties, such as moral goodness or well-being, are especially language dependent and thus infectable by vaguenss. However, the Pareto rule for ranking outcomes (outcomes are equally good if Pareto indifferent; one is better if Pareto superior; otherwise they are incomparable) , which can be paired with any arbitrary outcome set and any arbitrary account of well-being, yields a kind of incomparability that has nothing to do with language. Our formalization should be able to capture incomparability in that sense -- because the whole project of welfarism is to start with the Pareto rule and then extend it.

incompleteness that results when a variety of perspectives are relevant to ordering some set, and we require convergence among the perspectives in order to conclude that one item is better.[16]

A third objection is more radical. Note that a quasiordering drops the requirements of completeness but retains the standard transitivity properties associated with "at least as good as", "better than," and "equally good as." The more radical suggestion is that we drop transitivity. Imagine that an outcome set **O** is ordered by a relation $\succcurlyeq^{M^*}$ which is reflexive but not transitive, and that we define $\succ^{M^*}$ (intransitive betterness) and $=^{M^*}$ (intransitive equal goodness) in terms of $\succcurlyeq^{M^*}$. Then, at least in the case of choice under certainty, the following would seem to be a thinkable rule for choice: choose any action *a* such that its corresponding outcome *x* is not worse than any outcome, i.e., there is no *y* such that $y \succ^{M^*} x$.

However, I do not pursue the approach of dropping transitivity here, for a variety of reasons: (1) As a matter of reflective equilibrium, the formal devices used to rank outcomes should track our intuitive concept of "goodness." I suggest that our deepest intuitions about "goodness" involve reflexivity (that each item is equally good as itself, and that no item is better than itself). Denying transitivity may be less counterintuitive than denying reflexivity, but it is certainly more counterintuitive than denying completeness. (2) Although individuals, in their actual choices, may fail transitivity, such examples do not make a normative case for dropping transitivity. Such a case would be made if there were strong examples where we intuitively judge that *x* better than *y* better than *z*, but not *x* better than *z* (e.g., because of a "cycle," namely that *z* is better than x)[17], but the existence of such examples is contestable. (3) Dropping transitivity means a large restriction of the inferential properties of goodness, which ceteris paribus is undesirable. Without transitivity, if *x* is better than *y*, we can still infer that *y* is not better than *x*; but information about the comparative goodness of *x* and *y*, and of *y* and *z*, now does not warrant inferences about the comparative goodness of *x* and *z*. (4) Our formal techniques for handling intransitivity are not nearly as well developed as those for handling complete orderings or incomplete quasiorderings, and presumably would be much more complicated than those. This is a real strike against intransitivity, insofar as tractability is one desideratum in a decision procedure. (5) Extending this approach to choice under uncertainty creates serious difficulties, because denying transitivity means denying a very persuasive axiom for choice under uncertainty, namely dominance.

A final objection is that it is unnecessary to construct a full quasiordering. Why not just partition the set of outcomes into two subsets -- a set of optimal outcomes, and a set of suboptimal outcomes -- without any further ranking of the outcomes within these two subsets? This is the approach taken, for example, in the literature on "fair" allocations. This approach is perhaps adequate in handling an action set; but it may well be insufficient, in the case of choice

[16] See Chapter 3.

[17] A cycle is sufficient but not necessary for intransitivity. An asymmetric binary relation can be intransitive but acyclic. The quasiordering precludes any form of intransitivity in "at least as good as," "better than," and "equally good as," whether cyclic or acyclic.

under uncertainty, in handling an outcome set so as to generate a partition of the action set. What if each choice, with nonzero probability, yields some optimal outcomes but also some suboptimal outcomes? If the outcome set is merely partitioned into optimal and suboptimal outcomes, how do we move from the outcome set to a partition of the choice set into optimal and suboptimal actions?

In any event, if the reader prefers to characterize the outcome set in this more minimal way, the tools presented here certainly allow her to do so. Given a quasiordering $\succcurlyeq^{M}$ of an outcome set, the set of optimal outcomes consist of all outcomes each of which is not worse than any other outcomes in the set; suboptimal outcomes are all others.

In short, I believe a strong case can be made for structuring a welfare consequentialist decision procedure around a possibly incomplete quasiordering of actions and outcomes, rather than insisting on completeness, modeling incomparability via different formal devices than the quasiordering, allowing intransitivity, or simply characterizing outcomes and actions as optimal versus suboptimal. [18]

***

Thus far, I have focused on the outcome and action sets and their quasiordering features, but have said nothing systematic about persons or well-being. How, more precisely, do these figure into the decision procedure?

I assume that there is a population of N persons; the decision is supposed to be an impartial function of their well-being. Because *moral* decisionmaking, arguably, should be unconcerned with national or temporal differences, it is natural to think that the N individuals are the entire population of the world, past, present and future. However, the SWF framework and the others discussed in this book can also be applied to some subset of the world's intertemporal population – for example, a British decisionmaker might apply it to all British citizens.

Each of the N persons is a human being who is determinately a person.[19] Further, I assume throughout the book that N is a finite number and that each member of the N individuals exists in each and every outcome in $\mathbf{O}$. (In other words, the person comes into existence in either the first period or a subsequent period, in every outcome, and lives for at least one

[18] I have not mentioned the idea of "feasibility," which figures in much economic analysis. In my set-up, feasibility considerations most naturally come into play in constructing the choice set, or in determining probabilities. (For example, some hypothetical course of action may be impossible, given technological factors; and what would happen, if some course of action were pursued, will depend on technology.). Feasibility may also be used to prune down the outcome set. If every action which would yield outcome x with nonzero probability is infeasible, outcome x might be dropped from the outcome set.

[19] Again, I do not discuss the treatment of non-humans who are determinate persons, nor humans who are indeterminate persons or determinately non-persons, nor non-humans who are indeterminate persons. Further, it is not necessary for me to take a position regarding whether a human being is a person in virtue of her psychological properties, in virtue of having a soul, or for some other reason (although I find the psychological account most plausible.).

period.)[20] Relaxing these assumptions generates a number of very difficult problems in welfarist theory – nonidentity problems; the "repugnant conclusion"; infinite populations– that are briefly reviewed in the final chapter, but that this book does not attempt to resolve. It will be enough work to craft a decision procedure for the core case of a fixed, finite population.

The welfarist decision procedure will include some account of individual well-being. What is human well-being? That question is hotly debated by philosophers, economists, and psychologists, and will be discussed at length in Chapter 3. Does well-being consist in actual or idealized preference-satisfaction? In the occurrence of positive mental states, e.g., pleasant sensations? In various objective goods? The generic structure I set up, here, is meant to accommodate all these possibilities. The decision procedure includes *some* account of well-being, which may revolve around mental states, objective goods, preference satisfaction, or some combination.

This account of well-being has two critical functions. First, it helps generate the initial construction of the outcome set **O**. Remember that outcomes are incomplete characterizations of the different possible worlds that might result from choice. How to construct such characterizations – whether to describe outcomes in terms of consumption, or health, or public goods, or leisure, or happiness, or something else – is a matter of balancing the decision costs of reasoning with more finely specified outcomes, against the benefits of taking account of additional well-being relevant characteristics. What types of attributes of possible worlds are well-being relevant is, of course, not uncontroversial, but rather one of the critical questions that an account of well-being helps us to answer.

The second critical function of an account of well-being – and one that will be a central focus of the book -- is to help the decisionmaker arrive at a moral quasi-ordering of a given outcome set, **O**, once the outcome set has been specified. How does the account of well-being do this? It does so by taking the set **H** of life-histories; constructing various rankings on that set; and using this well-being information regarding the life-history set to help determine the moral quasi-ordering of the outcome set **O**.

A life-history simply means being some individual in some outcome. Formally, a life-history is simply a pairing of some individual in the set of N individuals, and some outcome in the outcome set **O**. The pairing (x; i) is one life history - - meaning being individual i in outcome x. The pairing (y; j) is another life history – meaning being individual j in outcome y. The life-history set **H** is the set of all life-histories, consisting of all such possible pairings. If **O** is finite (which it need not be), with O members, then the life history set has $N \times O$ elements; otherwise it is infinite.

[20] I am agnostic on the connection between the duration of each human person's life and the duration of the human being's life with which the person is associated. See Chapter 5.

The account of well-being furnishes principles whereby the decisionmaker can construct a quasiordering of the life-history set (call it "$\succcurlyeq^{WB}$"). The account of well-being may also furnish principles whereby the decisionmaker can construct other types of rankings associated with the life-history set, to be mentioned below. The idea of rankings of a life-history set may sound weird, but in fact it is very general and (I think) elegant. It accommodates the full range of accounts of well-being; crisply engages the topic of interpersonal comparisons; allows but does not require completeness; and leads to a generalized definition of the Pareto principles and a generalized characterization of welfarism.

The quasi-ordering of the life-history set **H** is a reflexive, transitive, possibly incomplete binary relation on **H**. $(x; i) \succcurlyeq^{WB} (y; j)$ means: life-history $(x; i)$ is at least as good for well-being as life-history $(y; j)$ or, to say the same thing more colloquially, individual i's well-being in outcome x is at least as great as individual j's well-being in outcome y. The well-being quasi-ordering makes an *intrapersonal* comparison when it takes two life histories involving the same person, $(x; i)$ and $(y; i)$, and says $(x; i) \succcurlyeq^{WB} (y; i)$. The well-being quasiordering makes an *interpersonal* comparison when it takes two life histories involving different persons, $(x; i)$ and $(y; j)$, i and j distinct, and says: $(x; i) \succcurlyeq^{WB} (y; j)$. We can now straightforwardly say what it means for an account of well-being to involve or allow for interpersonal comparisons (more precisely, interpersonal comparisons of well-being levels). That is just to say that there is *some* pair of life histories involving distinct individuals $(x; i)$ and $(y; j)$ -- at least one such pair, perhaps many more – such that $(x; i) \succcurlyeq^{WB} (y; j)$ according to the well-being quasiordering generated by that account.

It is of course a hotly disputed question whether the most attractive account of well-being allow for interpersonal comparisons. The notion of a well-being quasiordering over life-histories provides a formal structure for representing views on both sides of this dispute.

Consider the traditional, economic account of well-being, which reduces well-being to preference-satisfaction. This account of course allows for *intra*personal comparisons of well-being levels, saying specifically that $(x; i) \succcurlyeq^{WB} (y; i)$ iff individual i either prefers outcome x to outcome y or is indifferent. However, the preference satisfaction account does not allow for interpersonal comparisons: it never compares the well-being of one person in one outcome to the well-being of a different person in some outcome. By contrast, a mental-state account would make the comparison of life histories a function of individuals' mental states in the outcomes, and might make some comparisons of the form $(x; i) \succcurlyeq^{WB} (y; j)$, i and j distinct. For example, it might say that i's well-being in x is greater than j's well-being in y because i in outcome x experiences more pleasure and less pain than does j in outcome y.

Although the notion of a quasi-ordering of life-histories is very general with respect to both the range of accounts of well-being, and the question of interpersonal comparisons, it might be complained that my set-up is not general enough because it assumes that the kind of well-being which will be morally relevant is *lifetime* well-being. A life-history is an entire lifetime of

an individual in an outcome. To say that $(x; i) \succcurlyeq^{WB} (y; j)$ is to say that i's lifetime well-being in x is at least as great as j's lifetime well-being in y. What if the welfarist believes that the moral quasiordering of outcomes should be a function of "sublifetime well-being": individual well-being during some "time slice," such as annual or momentary well-being?

Because personal identity normally extends over time, the sublifetime view is problematic – a point more fully argued in Chapter 5. I could set up the welfarist structure in a yet more generic fashion, allowing for both a lifetime approach and a sublifetime approach. But that would yet further complexify an already complicated structure. Thus, anticipating the discussion in Chapter 5, my presentation here of a general format for welfare-consequentialist decisionmaking builds in a whole-lifetime approach.[21]

I have mentioned that an account of well-being may be used to construct various rankings associated with the life-history set. I have described one: the quasiordering of **H** ($\succcurlyeq^{WB}$). At a minimum, any account of well-being will generate such a ranking (with some degree of intra- and perhaps also interpersonal comparability of well-being levels). However, an account of well-being may also generate a quasiordering of *differences* between life-histories (call it "$\succcurlyeq^{D}$"), with some degree of intra- and perhaps interpersonal comparability of well-being differences. [22] A further possibility: an account of well-being might identify whether a given life-history is better, worse, equally good as, or incomparable with nonexistence. (As it turns out, the account of well-being I argue for in Chapter 3 will generate all three of these types of rankings.)

In turn, all the information about well-being contained in the various well-being rankings associated with **H** determines the moral quasi-ordering of the outcome set **O**. How exactly? That is one of the chief topics of controversy among welfarists: between welfarists who reject interpersonal comparisons, and welfarists who allow such comparisons; and, within the latter camp, between welfarists who disagree how interpersonally comparable utilities should drive the ranking of outcomes (in particular, between utilitarians who believe that utilities should be summed, and non-utilitarians who believe that the ranking should be sensitive to the distribution of utility).

At a minimum, however, all welfarists – because welfarism is a moral view – agree that the ranking of outcomes should be *impartial*. What exactly this means in the case of an account

---

[21] We could also see each pairing of an individual and an attribute as a separate "history," and specify welfarism in these terms. This possibility, like the sublifetime approach, is discussed and rejected in Chapter 5.

[22] Let us say that a difference quasiordering of life-histories, $\succcurlyeq^{D}$, ranks *pairs* of life-histories. It takes the form $[(x; i), (y; j)] \succcurlyeq^{D} [(z; k) , (w; l)]$, meaning that the difference between the pair of life-histories $(x; i)$ and $(y; j)$ is at least as great as the difference between the pair of life-histories $(z; k)$ and $(w; l)$. This idea is based on the literature on difference measurement. Formally, $\succcurlyeq^{D}$ is a binary relation on the product set $\mathbf{H} \times \mathbf{H}$, such that: (1) This relation is transitive and reflexive. (2) If $[(x; i), (y; j)] \succcurlyeq^{D} [(z; k), (w; l)]$, then $[(w; l), (z; k)] \succcurlyeq^{D} [(y; j), (x; i)]$. (3) If $[(x; i), (y; j)] \succcurlyeq^{D} [(x'; i'), (y'; j')]$, and $[(y; j), (z; k)] \succcurlyeq^{D} [(y'; j'), (z'; k')]$, then $[(x; i), (z; k)] \succcurlyeq^{D} [(x '; i '), (z '; k ')]$.
We can say that a difference quasi-ordering of life-histories *allows for interpersonal comparisons of well-being differences* iff there is at least one pair of pairs such that $[(x; i), (y; j)] \succcurlyeq^{D} [(z; k) , (w; l)]$, where it is not the case that $i=j=k=l$.

of well-being that does not allow for interpersonal comparisons is not obvious. Of course, the ranking should not depend on individuals' proper names, but what more to say is not clear. In the case of an account of well-being that *does* allow for interpersonal comparisons, impartiality is more straightforwardly characterizable. It can be expressed, in this case, as an "anonymity" (or "permutation") axiom: Given some outcome x, if the arrangement of well-being levels in y is simply a permutation of the well-being levels in x, then y and x are equally morally good. For a given pattern of well-being, it should not matter which particular person is at which level.[23]

Just as fundamentally, all welfarists agree that the ranking of outcomes should satisfy the Pareto indifference principle. And virtually all agree that this ranking should satisfy the principle of Pareto superiority (strong Pareto).

What are these principles? Standardly, within economics, the Pareto principle is formulated in terms of preferences. The Pareto indifference principle (in terms of preferences) says that: if each individual is indifferent between outcome x and outcome y, then the two outcomes are equally good. The strong Pareto principle (in terms of preferences) says that: if there is at least one individual who prefers outcome x to outcome y, and everyone else either prefers x to y or is indifferent, then x is better than y. The weak Pareto principle (in terms of preferences) says that: if everyone prefers outcome x to outcome y, then x is better than y. Note that the strong Pareto principle implies the weak principle, but not vice versa.

Because my account of welfarist decisionmaking is meant to allow for both preference-based and competing accounts of well-being, it needs to formulate the Pareto principles in a more general manner. This more general formulation is what John Broome calls "the principle of personal good." The Pareto indifference principle (general version) says: if there are two outcomes x, y, such that each person is just as well off in both outcomes, then x and y are equally morally good. Formally: if $(x; i) =^{WB} (y; i)$ for each individual i, then $x =^{M} y$. The strong Pareto principle (general version) says: if there are two outcomes x, y, such that at least one person is better off in x than y, and everyone is at least as well off in x, then x is morally better than y. Formally: if $(x; i) \succ^{WB} (y; i)$ for at least one individual, and $(x; i) \succcurlyeq^{WB} (y; i)$ for everyone, then $x \succ^{M} y$. The weak Pareto principle (general version) says: if everyone is better off in x than y, then x is morally better than y. Formally: if $(x; i) \succ^{WB} (y; i)$ for all individuals, then $x \succ^{M} y$. Note that the strong Pareto principle (general version) implies the weak Pareto principle (general version), but not vice versa.

Henceforth, I refer to these principles as the Pareto principles – emphasizing, again, that they are meant to reflect *whatever* account of well-being the welfarist uses (preferentialist or not) and whatever particular quasi-ordering over life-histories that account generates. I realize that

[23] Formally, a permutation is a mapping which is a bijection (one-to-one and onto). Assume that $\pi(.)$ is any permutation on the set of N individuals. Then anonymity says: if there is some $\pi(.)$ such that $(x; i) =^{WB} (y; \pi(i))$ for all i, then $x =^{M} y$. Difficulties arise in using the notion of a permutation to express an impartiality requirement where N is infinite, but not in the case under discussion in this book, with N finite.

there is some risk of terminological confusion, but the careful reader will not, I hope, be confused, and this terminology underscores the way in which my principles are faithful to, but generalize, the traditional economic formulation.

I assume that any welfarist principle for quasi-ordering an outcome set, whatever it may be, will satisfy the principle of Pareto indifference and the strong Pareto principle (which I'll also refer to as the principle of Pareto superiority). Why assume this? The principle of Pareto indifference captures the essence of welfarism. Welfarism says, informally, that the ranking of outcomes is solely a function of individual well-being. Moral goodness *supervenes* on well-being. Unless there is a well-being difference between two outcomes, they must be equally morally good. But this is exactly what Pareto indifference stipulates. If x and y are not ranked as equally morally good, and yet each individual is just as well off in x as in y, then some fact other than a fact about well-being must be driving the ranking.

Weak Pareto and strong Pareto do not logically follow from the notion that moral goodness supervenes on well-being. It is logically possible to accept that notion, and Pareto indifference, but give a higher moral ranking to outcomes in which everyone is worse off. For example, it is possible and not morally unthinkable to say that some outcome y in which well-being is unequally distributed is worse than some outcome x in which there has been "leveling down," i.e., everyone's welfare is lower but there is perfect equality of welfare. Still, virtually all welfarists accept not merely weak Pareto, but strong Pareto. This is true, not just of economists (strong Pareto is one of the most basic building blocks of welfare economics) but also of the moral philosophers who are sympathetic to welfarism. If everyone is strictly better off (regardless of the pattern of their gains), how could the outcome, too, not be a better one, all things considered? And (since we are willing to embrace this principle regardless of the pattern of individual gains) why wouldn't it extend to the case in which everyone is at least as well off and some are strictly better off?

This book will therefore take the strong Pareto principle (Pareto superiority) and Pareto indifference as basic axioms, along with impartiality. A key question for subsequent chapters will be: Can we develop an impartial principle for quasiordering outcome sets that conforms to the axioms of Pareto indifference and strong Pareto, but also goes beyond these axioms? Such a principle takes some pairs of outcomes that are *Pareto-noncomparable* – outcomes that are not ranked as better, worse, or equally good by the Pareto principle – and ranks one outcome in the pair as better, worse, or equally good. A key argument in subsequent chapters will be that the SWF approach is the most attractive way to do this.

Formally, we are looking for a principle for generating a quasi-ordering of a given outcome set that *extends* the Pareto quasi-ordering. Let us say that $x \succ^{\text{Paerto}} y$ iff $(x; i) \succ^{\text{WB}} (y; i)$ for at least one individual and $(x; i) \succcurlyeq^{\text{WB}} (y; i)$ for everyone. And let us say that $x =^{\text{Pareto}} y$ iff $(x; i) =^{\text{WB}} (y; i)$ for all i. (In other words, $x \succ^{\text{Paerto}} y$ means that x is ranked better than y by virtue of the strong Pareto principle. And $x =^{\text{Pareto}} y$ means that x and y are ranked as equally good by

virtue of Pareto indifference.) Then we can use these relations to define the Pareto quasi-ordering: $x \succcurlyeq^{\text{Pareto}} y$ iff $x \succ^{\text{Paerto}} y$ or $x =^{\text{Paerto}} y$. x and y are Pareto-noncomparable if it is neither the case that $x \succcurlyeq^{\text{Pareto}} y$ nor the case that $y \succcurlyeq^{\text{Pareto}} x$. Pareto-noncomparability will occur, for example, if some individuals are better off in x, but others are better off in y; and it can occur in other ways as well.

We want our principle $\succcurlyeq^{\text{M}}$ to be consistent with the Pareto quasiordering, meaning more precisely that (1) if $x \succ^{\text{Paerto}} y$, then $x \succ^{\text{M}} y$, and (2) if $x =^{\text{Pareto}} y$ than $x =^{\text{M}} y$. But, intuitively, an attractive principle for quasi-ordering an outcome set should go beyond the Pareto quasiordering. It will at least sometimes be the case that neither $x \succcurlyeq^{\text{Pareto}}$ y nor $y \succcurlyeq^{\text{Pareto}} x$, and yet $x \succcurlyeq^{\text{M}} y$.

What is wrong with just using $\succcurlyeq^{\text{Pareto}}$ as $\succcurlyeq^{\text{M}}$? Why not say that the moral quasi-ordering of outcomes just is the Pareto quasiordering? The problem is not that the Pareto quasi-ordering is *incomplete*. After all, the most attractive $\succcurlyeq^{\text{M}}$ may be incomplete as well. Rather, it is that the Pareto quasi-ordering is too incomplete. Intuitively, there are surely some pairs of outcomes that are rankable as better, worse, or equally good, even though some are better off in one outcome and others in the other.

This intuition is certainly shared by welfarists who believe in interpersonal comparisons. Imagine that a few very well off individuals are slightly better off in x than y, and many badly off individuals are substantially worse off in x than y. Surely we can say that x, on balance, is a morally a worse outcome than y.

But even welfarists who reject interpersonal comparisons intuit that some Pareto-noncomparable outcomes can be morally ranked. For example, the Kaldor-Hicks principle was spearheaded by economists who eschewed interpersonal comparisons but felt that the Pareto principle was too limited. The egalitarian-equivalent ordering, too, which we'll discuss in the next chapter, is an attempt to move beyond the Pareto quasiordering without making interpersonal comparisons.

In short: how shall we preserve Pareto indifference and Pareto superiority, but also rank at least some Pareto-noncomparable outcomes in an attractive way? This is the question to which we now turn.

Chapter Two: The SWF Approach and its Competitors

Chapter 1 outlined the agenda for the book: to elaborate a *welfarist* framework for morally evaluating governmental policies and other large-scale choices, one that has the basic welfarist features of being exclusively person-centered, consequentialist, and focused on well-being.

Chapter 1 also set forth the formal architecture of a welfarist choice-evaluation framework. It derives a moral ranking of an action set from a moral ranking of an outcome set. The ranking of the outcome set, in turn, is derived from information about individual well-being, encapsulated in a ranking of individual life-histories and other rankings associated with the life-history set. All these rankings are "quasiorderings," possibly incomplete.

The moral ranking of the outcome set must satisfy some basic criteria. First, it must satisfy the basic reflexivity and transitivity properties of any quasiordering: each outcome must be morally "at least as good as" itself, and if one outcome is morally at least as good as a second, in turn morally "at least as good as" a third, then the first must be morally at least as good as the third. Second, it must be "Pareto respecting," i.e., satisfy the two Pareto principles: Pareto indifference (if each individual in just as well off in one outcome as a second, then the two outcomes are equally morally good); and Pareto superiority[24] (if one or more individuals are better off in one outcome than the second, and everyone is at least as well off in that outcome, then the first outcome is morally better). Finally, it must be impartial.[25]

However, an attractive welfarist decision procedure will do more than satisfy the minimal criteria just articulated: producing an impartial quasi-ordering that conforms to the two Pareto principles. The Pareto quasi-ordering does all this, and yet is commonly seen to be too incomplete. There are surely some pairs of Pareto-noncomparable outcomes such that one can be ranked as morally better or worse than the other. Thus, we are looking here for a welfarist decision procedure that *extends* the Pareto quasi-ordering: that not only satisfies the minimal criteria just stated, but also ranks at least some Pareto-noncomparable outcomes – and does so in an "attractive" manner, one that we can accept in reflective equilibrium.

One way to structure a decision procedure in light of these aims is the SWF approach. This approach, originating in scholarship by Abram Bergson and Paul Samuelson in the 1930s, and developed most fully by welfare economists beginning in the 1970s, has the following

[24] This is formally the "strong Pareto" criterion, but I refer to it as "Pareto superiority" which is more colloquial. The reader is reminded that the Pareto indifference and Pareto superiority principles, as discussed throughout this book, are *generic* principles, formulated in terms of an individual's well-being rather than her preference satisfaction, so as to encompass the full range of theories of well-being.

[25] A criterion for ordering outcomes, at a minimum, does not yet use individuals' proper names or other expressions that refer to particular individuals ("you", "I"), in ranking outcomes. Given interpersonal comparability, a stronger impartiality criterion can be formulated, namely anonymity. See Chapter 1.

general features: it allows for interpersonal as well as intrapersonal welfare comparisons; uses "utility functions" to represent the ranking of life-histories, and to transform each outcome into a vector of utilities (or a set of such vectors); and ranks outcomes as a function of their corresponding utility vectors.

The chapter begins by introducing the SWF approach. This section is certainly *not* meant as a full elaboration or defense of the approach. That will occur in subsequent chapters. In particular, it will be the burden of Chapter 3 to develop an account of well-being that meets the needs of the SWF framework: an account that indeed allows for interpersonal comparisons and for the measurement of well-being via utility functions. Chapters 4 through 7 will grapple with a variety of other important questions that arise concerning the elaboration and practical implementation of the approach. Rather, this section of the chapter is meant to provide the reader an initial sense of how the framework works, and of why it holds promise as a way to structure a welfarist decision in a manner that not only meets the minimal criteria but also extends the Pareto quasi-ordering in an attractive way.

The bulk of the chapter will be devoted, not to discussing the SWF framework, but to criticizing the competing approaches to policy analysis that are currently dominant. Although SWFs are in fact used by some scholars for purposes of evaluating actual government policies – particularly within the scholarly field of optimal tax policy -- other approaches are currently much more widespread, both among scholars and in governmental practice. The dominant such frameworks are: cost-benefit analysis; inequality metrics, such as the well-known "Gini coefficient"; other equity metrics, in particular poverty metrics, social-gradient metrics, and tax incidence metrics; and QALY-based cost-effectiveness analysis. [26]

This chapter will critically examine these various frameworks, familiar to anyone who works on policy analysis. It will do so from the perspective of welfarism. The main thrust of this chapter will be to argue the following: each framework either fails to furnish an attractive

[26] Why aren't "happiness" metrics on the list of current policy evaluation frameworks? Quantifying individuals' happiness or "subjective well-being," via surveys, and studying the correlates of happiness, has become a huge topic in economics and psychology; and a number of prominent scholars in the field have suggested that governmental policy should be focused on the promotion of happiness. However, this suggestion has not – as yet – given rise to a well-developed policy-evaluation framework distinct from those already mentioned. Rather, insofar as they have attempted to quantify policy impacts, happiness scholars have mainly worked within existing frameworks. For example, some happiness scholars have calculated willingness-to-pay/accept values, for purposes of cost-benefit analysis, by using happiness surveys rather than the preference data traditionally employed by CBA researchers. Work has also been done using inequality metrics to study the distribution of happiness, and using poverty metrics to determine the extent to which individuals are below a subjective-well-being threshold.

Happiness research, thus, is not given its own rubric in this chapter, but is discussed in connection with the various policy frameworks addressed here. It is also discussed at different junctures in subsequent chapters.

basis for constructing an impartial, Pareto-respecting, quasiordering of an outcome set, or achieves this goal only by functioning as a variation on the SWF approach.[27]

The reader might wonder why I have chosen to organize the book by focusing first on the competitors to the SWF approach, in this chapter, and only then undertaking an extended discussion of the approach. Isn't this backwards?

I believe that a critical discussion of existing policy analysis frameworks is helpful in *motivating* a full-blown analysis of the SWF approach. By seeing the inadequacies in currently dominant frameworks, we can see why it makes sense to devote the considerable intellectual effort required to work through the variety of technical, philosophical, and empirical challenges that arise in elaborating the SWF approach – the challenges that I discuss in subsequent chapters, and that other scholars in the SWF tradition have also spent much effort engaging. However, the reader who prefers to read Chapters 3 through 7 first, and then return to this chapter, can certainly do so.

It also bears emphasis that the criticisms of cost-benefit analysis, inequality metrics, other equity metrics, and cost-effectiveness analysis presented in this chapter do not generally presuppose the possibility of interpersonal comparisons. In particular, I will show that certain ways of using these frameworks to order outcome sets fail to meet basic welfarist criteria. These particular approaches fail to yield even a quasiordering of outcome sets, or run afoul of Pareto indifference and Pareto superiority. What is critical to understand is that even the skeptic about interpersonal comparisons -- if she is a welfarist – should not embrace a choice-evaluation framework that violates the basic Pareto criteria, or that doesn't serve to produce a well-behaved ranking of outcomes (a ranking that respects the basic reflexivity, symmetry, and transitivity properties of "better than" and "equally good as"). Moreover, even the skeptic about

[27] More precisely, I will argue that distributively weighted CBA and the application of an inequality metric to utility vectors can function as variations of the specific SWF framework advocated in this book (namely, ranking outcomes by applying a continuous prioritarian SWF to the utility vectors corresponding to those outcomes). Applying a poverty metric to utility vectors is also a variation on the SWF framework – but one that, I will argue, employs a less attractive SWF (specifically, an SWF with an absolute threshold). Further, one might see various other ways of employing inequality metrics (what I shall term the "income evaluation function" approach, the "attribute evaluation function" approach, and the "egalitarian equivalent" approach) as variations on the SWF approach. (See below, discussing various "correspondence results" which show how these methodologies "correspond" to an SWF.) If so, these are also less attractive, I believe, than using a continuous prioritarian SWF in conjunction with utility functions as constructed in this book.

. Using CBA without distributive weights to order an outcome set, as well as the uses of social gradient metrics, incidence analysis, and cost-effectiveness analysis that I consider below, are not variations on the SWF approach and are unattractive from a welfarist perspective. The trio of approaches to inequality metrics mentioned in the prior paragraph that do not explicitly employ utilities (income-evaluation function, attribute-evaluation function, egalitarian equivalent) need not be seen as a variations on the SWF approach, since one can employ these tools but reject interpersonal comparability. Thus understood, they are less attractive than the SWF framework.

In a few cases, I will concede that non-SWF approaches to ranking outcome sets are "fallbacks," i.e., conditionally attractive if the case for interpersonal comparability and the existence of utiltiy numbers representing well-being fails.

interpersonal comparisons may recognize that a proposed choice-evaluation framework, albeit consistent with minimal welfarist criteria, is substantively unattractive.

In some instances, however, I will concede that certain ways of using current policy frameworks to rank outcome sets do meet minimal welfarist criteria and, further, are *conditionally* attractive – conditional on the premise that well-being is not interpersonally comparable and measurable by utilities. These approaches are plausible "fallbacks," in the event that the argument for interpersonal comparability and the SWF approach that I attempt to mount in subsequent chapters is unpersuasive. But there are various other ways of using current frameworks to order outcome sets that should be seen as problematic independent of the success of my eventual argument for the SWF approach.

One objection to the strategy followed in this chapter – examining currently dominant policy-evaluation frameworks from a welfarist perspective – is that some of these frameworks, in particular QALY-based cost-effectiveness analysis and social gradient metrics, are not intended as welfarist methodologies. It is certainly true that these two approaches are often defended in non-welfarist terms; whether they actually succeed in providing an appealing non-welfarist framework for policy choice is a different question, one that I will not attempt to engage. This book works *within* welfarism, for reasons belabored in Chapter 1 – and it is important (if only for the sake of completeness) to see why the whole gamut of existing non-SWF frameworks for policy evaluation are problematic from the perspective of welfarism.

Another objection is that existing frameworks have some useful role other than as functioning as the outcome-ranking component of a moral decision procedure. In particular, it is often suggested that cost-benefit analysis has an important *legal* role: for example, that non-tax bodies should employ CBA, with redistribution handled through the tax system; or that the use of a cost-benefit test by governmental bodies will benefit everyone in the long run.

The difficult problem of optimal legal structures is discussed in Chapter 8. The main focus of the book, however, is designing a welfarist framework for engaging in the distinctive species of normative evaluation that I have referred to as "moral" – the species of normative evaluation that is characterized by a focus on persons, by impartiality between persons, and by a willingness to transcend existing social norms. Existing frameworks (to the extent they are not merely variations on the SWF approach) fail to meet the needs of *this* sort of normative deliberation. Whether the frameworks fulfill some other function is not a question I address in this chapter.

The SWF Approach: An Introduction

This section begin by characterizing the SWF framework and discussing why it seems a promising blueprint for a welfarist choice-evaluation procedure. This section then reviews the existing literature on SWFs, surveying the history, theory, and current scholarly uses of the "social welfare function."

What is an SWF?

What is the SWF framework for policy choice? I offer the following account. As I shall explain in the next subsection, this is not precisely what is meant by a "social welfare function" in existing scholarship, but rather a generalization of the current definition -- one that accommodates the possibility of a quasiordering.

Remember that *any* welfarist procedure will contain an account of well-being that furnishes a quasiordering of the life-history set **H** (and perhaps other rankings associated with the life-history set), and will use this information to generate a moral ranking of the outcome set **O**. The SWF framework, as I conceptualize it, accomplishes these tasks in a distinctive way. First, it represents the rankings of the life-history set using *utility numbers*. Second, it produces a moral ranking of the outcome set as a function of the utility numbers associated with outcomes. Third, it assumes some degree of interpersonal comparability of well-being.

I will discuss these features of the SWF approach in turn. At the same time, my discussion will show how the SWF format can easily satisfy the minimal criteria that any welfarist decision procedure should meet: producing a ranking of any given outcome set which is a quasiordering, which satisfies the Pareto principles, and which is impartial. And it will show how the approach promises to extend the Pareto quasiordering in an attractive manner (assuming interpersonal comparisons are indeed possible).

(1) The first, characteristic feature of the SWF framework is that it uses utility functions to *represent* individual well-being. Formally, the SWF framework contains a set **U** of utility functions. Each utility function $u(.)$ belonging to **U** maps a life-history onto a real number. So $u(.)$ takes the form $u(x; i)$. And **U** then represents the various rankings associated with the life-history set. Those rankings are *reflected* by the numerical information contained in **U**.

How can utility numbers mirror well-being rankings? Consider, first, the well-being quasiordering of life-histories. This is what I referred to, in the last chapter, as "$\succcurlyeq^{WB}$". This is a reflexive, transitive, binary relation on the set **H** of life-histories. It has the form: $(x; i) \succcurlyeq^{WB} (y; j)$. What this says, in English, is that life-history $(x; i)$ is at least as good for well-being as life-history $(y; j)$. It is *reflexive*: each life-history is at least as good as itself. It is *transitive*: If one life-history is at least as good for well-being as a second, and the second is at least as good for well-being as a third, then the first life-history is at least as good as the third. Further, this relation is a quasiordering, which can – but need not – be complete. To say that relation $\succcurlyeq^{WB}$ is complete means that, for every pair of life-histories, either the first is at least as good for well-being as the second, or the second is at least as good for well-being as the first, or both. If the relation $\succcurlyeq^{WB}$ is incomplete, then there is a pair or multiple pairs of life-histories such that the first life-history is not at least as good as the second, and the second is not at least as good as the first. These pairs are *incomparable*.

Remember, also, that the relation "at least as good as" can be used to define the relations of "better than" and "equally good as."

Consider, first, the case in which the well-being quasiordering of life-histories is complete. In that case (bracketing technical complications, which will be discussed in the margin below) it will be possible to use a *single* utility function to represent the ranking of life-histories. In particular, a single utility function could mirror the ranking of life-histories via a very natural "representational rule," which makes the "at least as good as" relation between life-histories isomorphic to the "greater than or equal to" relation between real numbers.[28] In other words, in the case of a complete quasiordering of life-histories, we could (bracketing technical complications) identify some $v(.)$ such that $v(x; i) \geq v(y; j)$ iff $(x; i)$ is at least as good for well-being as $(y; j)$.

For example, imagine that the outcome set contains three outcomes, x, y, z, and the population contains two individuals, i and j, yielding a life-history set with six elements. And imagine that they are completely ordered, as shown in the following table. In this particular case, we can *represent* $\succcurlyeq^{WB}$ via the following assignment of numbers: $v(x; i) = 20$; $v(y; j) = 12$; $v(y; i) = 12$; $v(z; i) = 9$; $v(z; j) = 7$; $v(x; j) = 7$.

	$(x; i)$	$(x; j)$	$(y; i)$	$(y; j)$	$(z; i)$	$(z; j)$
$(x; i)$	Yes	Yes	Yes	Yes	Yes	Yes
$(x; j)$	No	Yes	**No**	No	No	Yes
$(y; i)$	No	Yes	Yes	Yes	Yes	Yes
$(y; j)$	No	Yes	Yes	Yes	Yes	Yes
$(z; i)$	No	Yes	No	No	Yes	Yes
$(z; j)$	No	Yes	No	No	No	Yes

In this table, "yes" means that the life-history in the row stands in relation $\succcurlyeq^{WB}$ to the life-history in the column, i.e., is "at least as good as" the life-history in the column. For example, the bold **No** means that the life-history $(x; j)$ is not at least as good as the life-history $(y; i)$, i.e., that not $(x; j) \succcurlyeq^{WB} (y; i)$.

[28] There are actually a variety of "representational rules" that allow utility numbers to represent well-being. In the case of a complete ordering of life-histories, we could assign life-histories utility numbers such that the better life-history is assigned a higher number; or, we could use the less natural rule which assign the better life history a *lower* number. In the case of a complete ordering of differences between life-histories, we could identify a utility function $u(.)$ which represents that difference ordering via the following natural rule: The difference between life-history $(x; i)$ and $(y; j)$ is at least as great as the difference between life-history $(z; k)$ and life-history $(w; l)$ iff $u(x; i) - u(y; j) \geq u(z; k) - u(w; l)$. But consider that if we define $v(.) = e^{u(.)}$, then $v(.)$ represents the very same difference ordering via this less natural rule: the difference between life-history $(x; i)$ and $(y; j)$ is at least as great as the difference between life-history $(z; k)$ and life-history $(w; l)$ iff $v(x; i) / v(y; j) \geq v(z; k) / v(w; l)$.

However, it is almost invariably assumed in the literature on SWFs that utility numbers represent well-being levels and differences via the natural rules, and that will be my assumption throughout the book as well.

Inspection will show that the relation set forth by this table is indeed a complete quasiordering. It is reflexive, meaning that each life-history is at least as good as itself (this can be seen by seeing that the diagonal entries are all "Yes"). It is transitive, meaning that if one life-history is at least as good as a second, and the second is at least as good as a third, then the first is at least as good as the third. Finally, it is complete, meaning that for each pair of life-histories, either the first is at least as good as the second, or the second is at least as good as the first, or both.

The complete quasiordering in this particular table gives rise to various "betterness" and "equally good as" relations between the life-histories, which can be compactly summarized as follows: Life-history (x; i) is better than life-history (y; j), which in turn is equally good as life-history (y; i), which in turn is better than life-history (z; i), which in turn is better than life-history (z; j), which is equally good as life-history (x; j). If we assign (x; i) the largest number; (y; j) the second largest; (y; i) the same number as (y; j); (z; i) a lower number; (z; j) a yet lower number; and (x; j) the same number as (z; j), *then* these numbers will represent the quasiordering. For example, as stated in the text, we can say that $v(x; i) = 20$; $v(y; j) = 12$; $v(y; i) = 12$; $v(z; i) = 9$; $v(z; j) = 7$; $v(x; j) = 7$.

By assigning numbers this way, we ensure that one life-history will be better than another iff it is assigned a higher number, and that one life-history will be equally good as a second iff it is assigned the same number . That in turn implies that one life-history will be at least as good as a second iff the number assigned the first is at least as large as the number assigned the second. Inspecting the table will show that the numbers assigned by $v(.)$ here do indeed represent the quasiordering. Take any pair of a row life-history and a column life-history. Whenever the entry in the table is "Yes," the row life-history is assigned a $v(.)$ which is greater than or equal to the column life-history. Whenever it is "No," the row-life-history is assigned a smaller $v(.)$ value.

Of course, this is *not* the only $v(.)$ that represents this quasiordering. Any assignment of numbers which gives the largest number to (x; i), the second largest to (y; j), and so forth, will work equally well. For example, consider $v^*(.)$ such that $v^*(x; i) = 1900$; $v^*(y; j) = 555$; $v^*(y; i) = 555$; $v^*(z; i) = 3$; $v^*(z; j) = 2.75$; $v^*(x; j) = 2.75$

In the more general case where the quasi-ordering of life-histories is incomplete, it clearly is *not* possible to represent that quasi-ordering via a single utility function. [29] But (absent technical complications) it *will* be possible to represent the quasi-ordering via a whole *set* of utility functions. We can construct a set of utility functions and use the "zone of agreement" between them to mirror the ranking of life-histories. In other words, whenever one life-history is at least as good as a second, each and every utility function in **U** will assign the first life-history a number which is at least as large as the number it assigns to the second. Whenever one life-history is equally good as a second, each utility function in **U** will assign the two the very same number. And whenever one life-history is incomparable with a second, the utility functions will

[29] Any real number is equal to, greater than, or less than any other. But in the case of a quasiordering, there are four possibilities regarding any two life-histories: namely that the first is equally good as the second, better than the second, worse than the second, or incomparable with the second. Incomparability is symmetric: If one life-history is incomparable with a second, then the second will be incomparable with the first. So we can't represent this relation by assigning the first life-history a larger number than the second, since the relation of "greater than" between numbers is asymmetric. Nor, in general, can we represent it by assigning the two life-histories the same number, since incomparability doesn't have the transitivity properties of "equal to". If one life-history is incomparable with a second, and a second with a third, the first need not be incomparable with the third. If one life-history is incomparable with a second, and the second is better than a third, the first need not be better than the third.
The table immediately following provides a simple example which shows why an incomplete quasiordering can't be represented by a single utility function.

"disagree" about their ranking: some will assign the first life-history a number which is greater than or equal to the second, but others will assign the first life-history a lower number.

In short (absent technical complications) it will be possible to represent the well-being quasiordering of life-histories via the following set-valued rule: Life history (x; i) is at least as good for well-being as life-history (y; j) iff, for all $u(.)$ belonging to **U**, $u(x; i) \geq u(y; j)$.

To see how the set-valued rule can function to mirror an incomplete quasiordering of life-histories, consider the following example.

Imagine that outcomes characterize individuals as having two attributes a and b (health, leisure, consumption, happiness, etc.) . Each life-history is therefore a combination of the two attributes. Assume attributes have different numerical levels. Assume that life-histories are ordered using the following "attribute dominance" relation: One life-history is at least as good as a second iff the a level of the first life history is at least as large as the a level of the second, *and* the b level of the first life history is at least as large as the b level of the second. This in turn means that two life-histories are equally good just in case they have the same a and b levels; and that one life-history is better than a second just in case both of its attribute levels are at least as large, and one is larger. (For example, (4, 8) is better than (3, 6). It is also better than (3, 8). It is incomparable with (3,9).)

This relation between life-histories is a quasiordering. It *cannot* be represented by a single utility function. It can, however, be represented by the following set-valued rule. Let one utility function $v^a(.)$ be such that it assigns life-histories a number which depends solely on the a attribute, increasing as that attribute does; and a second utility function, $v^b(.)$, be such that it assigns life-histories a number which depends solely on the b attribute, increasing as that attribute does. Most simply, we can say that $v^a(a, b) = a$, and that $v^b(a, b) = b$. Then if **U** consists of $v^a(.)$ and $v^b(.)$, it will be the case that $(a, b) \succcurlyeq^{WB} (a^*, b^*)$ according to the dominance relation iff $u(a, b) \geq (a^*, b^*)$ for all $u(.)$ belonging to **U**, i.e., just in case $v^a(a; b) \geq v^a(a^*, b^*)$ *and* $v^b(a; b) \geq v^b(a^*, b^*)$.

For example, imagine that each attribute can have levels 1, 2, or 3. This yields 9 life-histories, the quasiordering of which can be represented in the following graph. Each life-history is better than all others "west" or "south" of it; worse than all others "east" or "north" of it; and incomparable with the rest. For example, the life history with attributes (3,1) is better than (2,1),and (1,1); it is worse than (3,2) and (3,3); it is equally good as itself; and it is incomparable with (2, 2) , (2,3), (1,2), and (1,3). [chart]

Note that no single $v(.)$ can represent these relations. We clearly can't represent incomparability by saying that one life-history is incomparable with a second if it is assigned a larger number. The relation of incomparability is *symmetric*. As can be seen by the graph, if one life-history is incomparable with a second, the second is incomparable with the first. Nor can we represent incomparability by assigning two life-histories the very same number. That would imply that, if one life-history is incomparable with a second, and a second with the third, then the first is incomparable with the third. But incomparability is *intransitive*. For example, life-history (1,2) is incomparable with life history (3,1); life history (3,1) is incomparable with life-history (2,2); but life-history (1,2) is worse than life-history (2,2), not incomparable with it.

However, the relations of the life-histories can be represented via a **U** which includes both $v^a(.)$ which assigns each life-history the level of the a attribute, i.e., $v^a(1,1) = 1$, $v^a(1,2) = 1$, $v^a(1,3) = 1$, $v^a(2,1) = 2$, etc.; and a $v^b(.)$ which assigns each life-history the level of the $b(.)$ attribute. It can be seen that one-life history is at least as good as second iff $v^a(.)$ and $v^b(.)$ assign the first life-history a number at least as large as the second.

The reader may wonder about the typicality of the example. Why think that the set-valued rule is *generally* a good recipe for representing an incomplete quasiordering of life-histories?

Imagine, first, that we *start* with a set **U** of utility functions, which have been identified in some manner, and we then *construct* a relation $\succcurlyeq^{WB}$. We use the set-valued rule to *define* $\succcurlyeq^{WB}$. We say that $(x; i) \succcurlyeq^{WB} (y; j)$ iff for all $u(.)$ belonging to **U**, $u(x; i) \geq u(y; j)$, and that $(x; i)$ and $(y; j)$ are incomparable iff (1) it is not the case that $(x; i) \succcurlyeq^{WB} (y; j)$ for all $u(.)$ belonging to **U**, and (2) it is not the case that $(y; j) \succcurlyeq^{WB} (x; i)$ for all $u(.)$ belonging to **U**. The reader can readily ascertain that a relation $\succcurlyeq^{WB}$ constructed in this manner will be a quasiordering: it will be reflexive and transitive.[30]

Indeed, this constructive strategy is more or less the strategy that I will pursue in chapter 3. Very roughly, the idea will be this: I will start with each individual's idealized extended preferences over life-histories, which I will assume to be complete; identify a utility function that represents these preferences; pool these utility functions to create **U**; and then define $\succcurlyeq^{WB}$ by saying that life history $(x; i)$ is at least as good for well-being as life-history $(y; j)$ iff, for all $u(.)$ belonging to **U**, $u(x; i) \geq u(y; j)$. [31]

But what if a decision procedure *starts* with a well-being quasiordering of life-histories as a primitive? Why think, in that case, that the quasiordering can necessarily be represented via the set-valued rule? This is where the technical complications arise. For reasons discussed in the margin, there is good reason to think that a quasiordering of life-histories will *often* (if not always) be representable using the set-valued rule. And, indeed, a set-valued approach to representing quasiorderings is a pervasive technique in scholarship on quasiorderings. [32]

---

[30] To see why a $\succcurlyeq^{WB}$ relation thus constructed is *reflexive*, take any $(x; i)$. Consider that, for all $u(.)$ belonging to **U**, $u(x; i) = u(x; i)$. Therefore, for all $u(.)$ belonging to **U**, $u(x; i) \geq u(x; i)$. Because $\succcurlyeq^{WB}$ has been defined such that one life-history $\succcurlyeq^{WB}$ a second if and only if its $u(.)$ value is at least as high for all $u(.)$ in **U**, it follows that $(x; i) \succcurlyeq^{WB} (x; i)$.

To see why a $\succcurlyeq^{WB}$ thus constructed is *transitive*, consider any triple of life-histories such that $(x; i) \succcurlyeq^{WB} (y; j)$ and $(y; j) \succcurlyeq^{WB} (z; k)$. Because $\succcurlyeq^{WB}$ has been defined such that one life-history $\succcurlyeq^{WB}$ a second if and only if its $u(.)$ value is at least as high for all $u(.)$ in **U**, it follows that $u(x; i) \geq u(y; j)$ for all $u(.)$ belonging to **U**; and that $u(y; j) \geq u(z; k)$ for all $u(.)$ belonging to **U**. It thus follows, mathematically, that $u(x; i) \geq u(z; k)$ for all $u(.)$ belonging to **U**. Because $\succcurlyeq^{WB}$ has been defined such that one life-history $\succcurlyeq^{WB}$ a second if and only if its $u(.)$ value is at least as high for all $u(.)$ in **U**, it follows – in turn – that $(x; i) \succcurlyeq^{WB} (z; k)$.

[31] This is only a rough statement of the approach in Chapter 3, because I actually employ utility functions that represent each individual's extended preferences over life-history lotteries, and comparisons to nonexistence. Further, each individual's extended preferences are actually represented by an entire individual set of utility functions; these are then pooled to form **U**.

[32] A lovely result in the theory of orderings is that *every* quasiordering can be represented by a set of complete orderings. For this reason, a set-valued approach to representing quasiorderings is the standard approach used in the formal literature. It is not the case that every complete quasiordering can, in turn, be represented by a single utility

The reader may have a different worry. In defending my use of the concept of a quasiordering in formalizing the structure of a welfarist choice-evaluation procedure, I stressed its flexibility: a quasiordering *can* be incomplete, but need not be. If the reader believes that completeness is a critical property of some ranking employed in the procedure (the ranking of life-histories, outcomes, actions, etc.), she can stipulate that this quasiordering must be complete. But here's the worry: why is the set-valued rule appropriate for representing the well-being quasiordering of life-histories in the case where that is complete?

The answer to that question is simple: If the utility functions in **U** bear an appropriate relation to each other, the ranking of life-histories achieved by the set-valued rule will be complete. In particular, if individuals have identical idealized extended preferences over life-histories, the methodology for constructing **U** employed in Chapter 3 will yield a set **U** whose elements are unique "up to a positive ratio transformation." In other words, **U** consists of a $u(.)$ and all positive multiples thereof. If $u(.)$ and $v(.)$ are both in **U**, and $u(x; i)$ has a certain value, then $v(.)$ will have the value $rv(x; i)$, where r is some positive constant. If indeed the elements of **U** are unique up to a positive ratio transformation, then it is straightforward to see that the set-valued rule yields a complete ordering of **H**.[33]

What about the other rankings that may be associated with **H**? I assume that any welfarist decision procedure at least include the relation $\succcurlyeq^{WB}$. But it may also include other rankings associated with the life-history set. A welfarist decision procedure that does include these additional rankings, and employs the SWF format, will also represent *these* rankings using utility numbers.

One such ranking is the quasiordering of *differences* between life-histories. The natural representational idea, here, is that numerical differences between utility numbers can be employed to mirror well-being differences between lives. That idea – generalized to allow for incomparability – yields the following rule for representing the difference quasiordering: The

---

function. That *will* be the case if the set of items being ordered is finite or countably infinite. It *may* but need not be the case if that set is uncountably infinite.

[33] Consider the case where **U** consist of a single $u(.)$ and all positive multiples thereof. We are using **U** to represent a quasiordering of life-histories, via the basic representational rule which says that one life-history is at least as good as a second iff its $u(.)$ value is at least as large, for all $u(.)$ in **U**. If we take any $u^*(.)$ in **U**, and say that one life-history is at least as good as a second according to $u^*(.)$, that quasiordering is obviously complete. But if we take any other $u^{**}(.)$ from **U**, it will order life-histories exactly the same way as $u^*(.)$, since both are positive multiples of $u(.)$. Assume that $u^*(.) = r^*u(.)$ and $u^{**}(.) = r^{**}u(.)$, with both r^* and r^{**} positive. Then whenever $r^*u(x; i) \geq r^*u(y; j)$, it will also be the case that $r^{**}u(x; i) \geq r^{**}u(y; j)$.

difference between (x; i) and (y; j) is at least as great as the difference between (z; k) and (w; l) iff, for all $u(.)$ belonging to **U**, $u(x; i) - u(y; j) \geq u(z; k) - u(w; l)$.[34]

Another such ranking is the comparison of life-histories to nonexistence, naturally represented as follows: life-history (x; i) is at least as good as nonexistence iff, for all $u(.)$ belonging to **U**, $u(x; i) \geq 0$.

(2) A second characteristic feature of the SWF format is that it ranks outcomes as a function of their associated utility numbers. The SWF format *uses* **U** to generate the moral quasiordering of the outcome set **O**.

How does this occur? Each utility function $u(.)$ has a dual aspect. To begin, each $u(.)$ is a "scalar valued" function from the elements of **H** onto the real numbers. In other words, a given $u(.)$ maps each life-history onto a single, real number. For example, $u(x; i) = 5$, or $u(y; j) = 6$.

But note that each outcome in **O** corresponds to a whole package of life-histories, with N elements (N being the population size). For example, outcome x corresponds to the package of life-histories (x; 1), (x; 2), …, (x; N), where (x; 1) means being individual 1 in outcome x, (x; 2) means being individual 2 in outcome x, and so forth. So each $u(.)$ can also be understood as a"vector valued" function, mapping the elements of **O** (outcomes) onto N-entry lists or "vectors" of utility numbers.

In general, a function that maps outcomes onto N-entry vectors has the form, $v(x) = (v_1(x), v_2(x), \ldots, v_N(x))$, where $v_1(x)$ is the first component of this N-entry vector, $v_2(x)$ the second component, ..., $v_i(x)$ the ith component, …, $v_N(x)$ the last component. So let us start with the utility functions in **U**, understood as scalar-valued functions from life-histories to real numbers. For each such scalar-valued $u(.)$, we can define its vector-valued counterpart as follows: $u(x)=(u(x; 1), u(x, 2), \ldots, u(x; i), \ldots, u(x; N))$. In other words, the first component of $u(x)$ is the utility assigned by scalar-valued $u(.)$ to the life-history of individual 1 in x; the second component of $u(x)$ is the utility assigned by scalar-valued $u(.)$ to the life-history of individual 2 in x; …the ith component of $u(x)$ is the utility assigned by scalar-valued $u(.)$ to the life-history of individual i in outcome x; and so forth.

For example, imagine that $N = 3$, and that a utility function $u(.)$ in **U** assigns the following numbers to life-histories: $u(x; 1) = 10$, $u(x; 2)=7$, $u(x; 3) = 8$. Then the vector-valued counterpart of $u(.)$ maps x onto the following three-entry vector: (10, 7, 8).

Henceforth, I will use the single symbol $u(.)$ to mean both a scalar-valued function from life-histories to single real numbers, and its vector-valued counterpart. The context will make clear which I'm referring to. Where "$u(.)$" is being used to mean a vector-valued function from

[34]Under what conditions can a difference quasiordering be represented by this rule? That is an interesting question, but not one I need to address, because my difference quasiordering is *constructed* from preexisting utility functions that represent individuals' extended preferences over life-history lotteries. See Chapter 3.

outcomes to *N*-entry vectors, I will often use the following symbolism: $u(x) = (u_1(x), u_2(x), \ldots, u_N(x))$. The number $u_i(x)$ is the *i*th component of this vector; it is the utility number assigned to individual *i* in outcome *x*, and is just equal to $u(x; i)$, the utility of having the life-history that consists in being individual *i* in outcome *x*.

So now we have the materials for moving *from* the set **U**, which numerically represents our account of well-being, *to* a moral quasiordering of the outcome set. Consider first the highly simplified case in which **U** has a single element *u*(.). In that case, *u*(.) maps each outcome in *x* onto a single utility vector, and a SWF will then rank outcomes as a function of their corresponding vectors. How will it do so? It might employ a mathematical function *w*(.), which maps *N*-entry vectors of numbers onto a single real number. One such *w*(.) might take the utilitarian form, namely $w(u(x))= u_1(x)+u_2(x)+\ldots+u_N(x)$ or, for short, $w(u(x)) = \sum_{i=1}^{N} u_i(x)$. In the highly simplified case at hand, the SWF format using the utilitarian *w*(.) would say: outcome *x* is morally at least as good as outcome *y* iff $\sum_{i=1}^{N} u_i(x) \geq \sum_{i=1}^{N} u_i(y)$ (meaning that the sum of individual utilities in *x* is at least as great as the sum of individual utilities in *y*). Another such *w*(.) might take the continuous prioritarian form of summing "transformed" utilities, namely $w(u(x))= g(u_1(x))+g(u_2(x))+\ldots+g(u_N(x))$, where *g*(.) is a strictly increasing and strictly concave function (more on what this means in a moment). This approach would say: outcome *x* is morally at least as good as outcome *y* iff $\sum_{i=1}^{N} g(u_i(x)) \geq \sum_{i=1}^{N} g(u_i(y))$ -- meaning that the sum of transformed utilities in *x* is at least as large as the sum of transformed utilities in *y*. Yet another such *w*(.) takes the "rank weighted" form of summing utilities multiplied by fixed weights, with larger utilities receiving smaller weights.[35]

Much of the policy-analysis work that employs SWFs *does* rank utility vectors using a mathematical function *w*(.). However, the theoretical literature on SWFs stresses an important point – that there may be plausible rules for ranking utility vectors which cannot be encapsulated in a mathematical function. The most important such example is the so-called "leximin" rule for ranking utility vectors.[36] So even in the case now under discussion, where **U** includes only a single element *u*(.), and each outcome corresponds to a single utility vector, a more general definition is needed. Let us say this: an SWF, in that case, ranks a given pair of outcomes – saying that the first is morally better than, worse than, equally good as, or incomparable with the second -- by applying some rule *R* to the utility vectors corresponding to those outcomes. The

[35] Formally, the rank-weighted rules uses *N* numbers $a_1, \ldots, a_N$, where each is larger than the next. For each utility vector, produce a corresponding rank-ordered vector by ordering its elements from smallest to largest, splitting ties arbitrarily. Then *w*(.) equals a_1 times the first element of the rank-ordered vector plus a_2 times the second element plus … plus a_N times the last element. See Chapter 4.

[36] The discussion in Chapter 4 will also cover other SWFs that cannot be represented by a mathematical function, namely a "sufficientist" SWF and a prioritarian SWF with an absolute threshold.

rule *R might* be: *x* is at least as good as *y* iff $w(u(x)) \geq w(u(y))$, where $w(.)$ is a mathematical function that maps vectors onto single, real numbers. However, *R* might have some other form.

The generalization of the idea of an SWF to the case where **U** is a non-singleton set of utility functions is now at hand. In that more general case, we can say this: For a given outcome set **O** and a given **U**, each $u(.)$ will map a given pair of outcomes, *x* and *y*, onto a pair of utility vectors. And there will be some rule *R** that morally ranks the two outcomes as a function of the *set* of pairs of utility vectors associated with the pair of outcomes, *x* and *y*, by all the elements in **U**. [37]

This is *quite* general and admittedly pretty abstract. To make things more tractable, I will assume that the SWF format – in the general case of an entire set **U** of utility functions -- takes what might be called the "supervaluationist" form. What does this mean? There is some rule *R* for ranking pairs of utility vectors, be it the utilitarian $w(.)$ just mentioned, the continuous prioritarian $w(.)$, the rank-weighted $w(.)$, the leximin rule, or some other rule *R*. What the SWF format will then do – I will assume – is rank outcomes using the following rule: *x* is morally at least as good as *y* iff, for all $u(.)$ belonging to **U**, the vector $u(x)$ is ranked at least as good as the vector $u(y)$ by rule *R*.[38]

In particular, the utilitarian SWF (in the general case of an entire set **U**) has the form: outcome *x* is morally at least as good as outcome *y* iff, for all $u(.)$ belonging to **U**, $\sum_{i=1}^{N} u_i(x) \geq \sum_{i=1}^{N} u_i(y)$. The continuous prioritarian SWF has the form: outcome *x* is morally at least as good as outcome *y* iff, for all $u(.)$ belonging to **U**, $\sum_{i=1}^{N} g(u_i(x)) \geq \sum_{i=1}^{N} g(u_i(y))$. (The particular version of a continuous prioritarian SWF I favor has the Atkinsonian form, namely: outcome *x* is morally at least as good as outcome *y* iff, for all $u(.)$ belonging to **U**, $\frac{1}{1-\gamma}\sum_{i=1}^{N} u_i(x)^{1-\gamma} \geq \frac{1}{1-\gamma}\sum_{i=1}^{N} u_i(y)^{1-\gamma}$.) The rank-weighted SWF has the form: outcome *x* is morally at least as good as outcome *y* iff, for all $u(.)$ belonging to **U**, the sum of rank-weighted utilities in $u(x)$ is at least as large as the sum of rank-weighted utilities in $u(y)$. The leximin SWF has the form: outcome *x* is morally at least as good as outcome *y* iff, for all $u(.)$ belonging to **U**, $u(x)$ is

[37] This formulation is a generalization of the "independence of irrelevant alternatives" condition used in the literature on social welfare functions. It rules out more complicated ways of using a set **U** to determine the ranking of outcomes: for example, allowing the ranking of a given pair of outcomes *x* and *y* to depend on the utilities assigned by the elements of **U** to all the outcomes in the outcome set, not just *x* and *y*; or by allowing that ranking to depend both on the numbers assigned to *x* and *y* by the elements of **U**, and on other information about the two outcomes.

[38] Rule *R* need *not* provide a complete ranking of utility vectors. The SWFs discussed to this point, and considered in Chapter 4, do in fact employ a rule *R* that yields a complete ranking of utility vectors. However, at the end of Chapter 4 I will consider the possibility that divergence in individuals' moral preferences yields a rule *R* which is not complete.

ranked at least as good as $u(y)$ using the leximin rule. As we will see in Chapter 4, there are yet further possibilities.

My assumption that the SWF format takes the "supervaluationist" form is not just a matter of tractability. It is also very easy to show that such an approach effortlessly satisfies the minimal welfarist criteria. Imagine that R has the following features: (a) R is anonymous, meaning that if one utility vector is just a reordering or "permutation" of a second, it ranks them as equally good; and (b) R is "Paretian," meaning that if every utility number in $u(x)$ is at least as large as the corresponding entry in $u(y)$, with at least one strictly larger, R ranks $u(x)$ as better than $u(y)$. If so, the SWF which incorporates R in supervaluationist fashion will necessarily satisfy the minimal criteria: it will produce a *quasiordering* of the outcome set **O**; this quasiordering will respect the principles of Pareto indifference and Pareto superiority; and it will be impartial. [39]

The SWFs mentioned in the prior paragraph are all Paretian and anonymous. So are the other SWFs discussed in Chapter 4. For the remainder of the book, unless otherwise noted, I will use the term "SWF" to mean a Paretian anonymous SWF. An SWF which lacks these characteristics will not reliably order outcome sets so as to meet minimal welfarist criteria, and does not merit consideration in this book.

(3) A third characteristic of the SWF format is that it assumes some degree of interpersonal comparability of well-being.

The reader will be reminded what this means. To begin, the literature on interpersonal comparisons emphasizes that we should differentiate between two different kinds of comparability, neither of which entails the other: interpersonal comparability of well-being *levels* and interpersonal comparability of well-being *differences*. Using the generic formal architecture of a welfarist decision procedure employed in this book, I have expressed interpersonal level comparability as follows: there is at least one pair of life-histories involving different subjects, such that the two are not incomparable according to the quasiordering of **H**. In other words, there exists $(x; i)$ and $(y: j)$, i and j distinct, such that $(x; i) \succcurlyeq^{WB} (y; j)$. And I have expressed interpersonal difference comparability in terms of the quasiordering of differences between life histories (assuming there is one): there exists at least one group of four life-histories, $(x; i)$, $(y; j)$, $(z; k)$, and $(w; l)$, such that the four subjects are not all identical, and such that the difference between the first two is at least as large as the difference between the second two.

Where the welfarist choice-evaluation procedure uses the SWF format – so that the rankings associated with **H** are represented by a set **U** of utilities – these two kinds of

[39] Because R is anonymous, the SWF is clearly impartial in the basic sense of not using proper names or other expressions that refer to particular individuals. If **U** is such as to allow for interpersonal level comparability -- which means that impartiality can be formulated in the stronger sense of giving an equal moral ranking to x and y if the well-being levels in y are a permutation of the levels in x -- the SWF which incorporates R in supervaluationist fashion will be impartial in this stronger sense.

interpersonal comparability can be translated into conditions on **U**. Interpersonal level comparability means: there exists at least one pair of life histories (x; i) and (y: j), i and j distinct, such that $u(x; i) \geq u(y; j)$ for all $u(.)$ belonging to **U**. Interpersonal difference comparability means: there exists at least one group of four life-histories, (x; i), (y; j), (z; k), and (w; l), such that the four subjects are not all identical, and such that $u(x; i) - u(y; j) \geq u(z; k) - u(w; l)$ for all $u(.)$ according to **U**.

The existing literature on social welfare functions assumes that the utility numbers which function as inputs to the social welfare function are indeed interpersonally comparable to some extent – either in terms of levels, or in terms of differences, or both. And my conceptualization of the SWF framework will incorporate this critical requirement.

The reader might wonder why the SWF framework needs interpersonal comparability. Imagine that the account of well-being is such that neither well-being levels, nor well-being differences, are interpersonally comparable *at all*. However, the account makes intrapersonal level comparisons, and also perhaps intrapersonal difference comparisons, and these are faithfully represented by a set **U**. What would be wrong with taking some rule R, such as the utilitarian, prioritarian, rank-weighted, or leximin SWF, and saying: x is at least as good as y iff, for all $u(.)$ belonging to **U**, $u(x)$ is at least as good as $u(y)$ according to R?

A full rigorous answer to this question is beyond the scope of this book. But for reasons discussed in the margin, there is a real worry that many plausible SWFs will "collapse" to the Pareto quasiordering absent some degree of interpersonal comparability. Remember that a key ambition of a welfarist decision procedure is to *extend* the Pareto quasiordering: to produce a ranking of outcomes that is consistent with Pareto indifference and Pareto superiority, but also ranks at least some pairs of Pareto-noncomparable outcomes in an attractive manner. Without some degree of interpersonal comparability, this ambition may well fail.[40]

---

[40] Imagine that $u^+(.)$ is in **U** and implies various intrapersonal comparisons of well-being levels and differences. Note that if $u^*(.)$ is constructed from $u^+(.)$ via individual-specific positive affine transformations, then $u^*(.)$ implies the very same intrapersonal level and difference comparisons as $u^+(.)$. In other words, $u^*(x; i) = a_i u^+(x; i) + b_i$, with a_i positive. If the account of well-being at hand allows for no *interpersonal* comparisons, then it would be arbitrary not to add $u^*(.)$ to **U**. (After all, it contains the very same intrapersonal information as $u^+(.)$). But the worry now is that if we have an SWF which ranks Pareto-noncomparable outcomes in a particular manner, we can eviscerate that ranking by adding $u^*(.)$ to **U**.

To see how this works for the utilitarian SWF, imagine that x and y are Pareto-noncomparable but the utilitarian SWF says that x is at least as good as y (rather than incomparable). In other words, for all $u(.)$ in **U**, $\sum_{i=1}^{N} u_i(x) \geq \sum_{i=1}^{N} u_i(y)$. Because x and y are Pareto-noncomparable, there must be at least one individual k and at least one $u^+(.)$ in **U** such that $u^+(y; k) > u^+(x; k)$. The idea, now, is that if we choose $u^*(.)$ appropriately, we can "blow up" the difference between (x; k) and (y; k), as calculated by $u^*(.)$, so that the utilitarian SWF no longer favors x. More precisely, for all individuals j distinct from k, set $K = \sum_j [u_i^+(x) - u_i^+(y)] > 0$. Now define $u^*(.)$ as follows. For all j distinct from k, $u^*(\_\_; j) = u^+(\_\_; j)$. In the case of k, pick a_i so that $a_i > K / [u^+(y; k) - u^+(x; k)]$. Set $u^*(\_\_; k) = a_i u^+(\_\_\_; k)$.

And what if we *can* make interpersonal comparisons? Given sufficient comparability, we will be able to extend the Pareto quasiordering. Consider first the simple, limiting case of the well-being account presented in Chapter 3: where **U** is unique up to a positive ratio transformation. In that case, there is full interpersonal and intrapersonal comparability of well-being levels and differences. If so, it is easy to see, all the standard SWFs mentioned here – the utilitarian SWF, leximin SWF, rank-weighted SWF, and prioritarian SWF (in the Atkinson form) – will produce a complete ordering of outcomes.[41]

However, it is also not hard to show that – even if **U** is not so well-behaved, and there is some incomparability in the ranking of life-histories or differences – these SWFs may well be able to extend the Pareto ordering.

> Consider the case in which life-histories have two attributes, which have a numerical level greater than zero. **U** is based upon two different utility functions. One utility function, $u(.)$, values life-histories by summing the attributes. That is, $u(a, b) = a+b$. A second utility function, $v(.)$, values them by multiplying

---

With $u^*(.)$ included in **U**, it is no longer the case that, for all $u(.)$ in **U**, $\sum_{i=1}^{N} u_i(x) \geq \sum_{i=1}^{N} u_i(y)$. Note that $\sum_{i=1}^{N}[u_i^*(y) - u_i^*(x)] = (u_k^*(y) - u_k^*(x)) + \sum_{j \neq k}[u_j^*(y) - u_j^*(x)]$. The second term is $-K$. The first term is a_i $[u^+(y; k) - u^+(x; k)]$ which, by construction, is a positive number greater than K.

A similar strategy can, I believe, be used for many other plausible SWFs – in particular, the SWFs mentioned in this section, i.e., the leximin, rank-weighted, and continuous prioritarian SWFs – to show that without interpersonal comparability these SWFs "collapse" to the Pareto quasiordering.

[41] In the case of the utilitarian SWF, this is very easy. **U** consists of $u(.)$ and all utility functions that are positive multiples. Take some $u^*(.)$ belonging to **U**, where $u^*(.) = r^*u(.)$ and r^* is positive. Consider any two outcomes x and y. Using the utilitarian SWF, the ranking of x and y according to $u^*(.)$ depends on summing their utilities. In other words, x is morally at least as good as y according to $u^*(.)$ just in case $u^*(x; 1) + u^*(x; 2) + \ldots + u^*(x; N) \geq u^*(y; 1) + u^*(y; 2) + \ldots + u^*(y; N)$. That relation is complete: for any two outcomes, x and y, and any given $u^*(.)$, either the first outcome's sum of utilities is greater than or equal to the second outcome's sum of utilities, or vice versa, or both (in the case of equality) . But note that, for any two utility functions in **U**, $u^*(.)$ and $u^{**}(.)$, the two will rank x and y the very same way using the utilitarian SWF, since $u^*(.) = r^*u(.)$ and $u^{**}(.) = r^{**}u(.)$, with both r^* and r^{**} positive. Therefore the ranking achieved by the utilitarian SWF -- using the formula x is at least as good as y iff $u(x; 1) + u(x; 2) + \ldots + u(x; N) \geq u(y; 1) + u(y; 2) + \ldots + u(y; N)$ for all $u(.)$ belonging to **U**-- is complete too.

Similar reasoning shows that the ranking of life-histories by the leximin and rank-weighted SWFs are complete. In the case of the continuous prioritarian SWF, with **U** unique up to a positive ratio transformation, the ranking will *not* necessarily be complete. This is because, in general, with $g(.)$ a strictly increasing and concave function, it is *not* necessarily the case that $g(u(x; i)) \geq g(u(y; j))$ iff $g(ru(x; i)) \geq g(ru(y; j))$. However, this *is* true in the case of the Atkinsonian SWF, where $g(u(x;i)) = \frac{1}{1-\gamma}u(x;i)^{1-\gamma}$. With **U** unique up to a positive ratio transformation, pick some $u^*(.) = r^*u(.)$, with r^* positive. With u^* in hand, the Atkinsonian SWF says that x is morally at least as good as y iff $\frac{1}{1-\gamma}\sum_{i=1}^{N}(r^*u_i(x))^{1-\gamma} \geq \frac{1}{1-\gamma}\sum_{i=1}^{N}(r^*u_i(y))^{1-\gamma}$. But that is true just in case $\frac{(r^*)^{1-\gamma}}{1-\gamma}\sum_{i=1}^{N}(u_i(x))^{1-\gamma} \geq \frac{(r^*)^{1-\gamma}}{1-\gamma}\sum_{i=1}^{N}(u_i(x))^{1-y}$. And that in turn will be true if r^* is replaced by r^{**}, with r^{**} also positive, thus for any other $u^{**}(.)$ in **U** equaling $r^{**}u(.)$.

the attributes. That is, $v(a, b) = ab$. **U** consists of $u(.)$ and all positive multiples (represented as "$ru(.)$"), plus $v(.)$ and all positive multiples (represented as "$sv(.)$").[42]

For any life-history, the zone of incomparability can be represented as follows. [Chart]

Given $x = ((a_1, b_1), (a_2, b_2) \ldots. (a_N, b_N))$ and $y= ((a_1^*, b_1^*), (a_2^*, b_2^*) \ldots. (a_N^*, b_N^*)$, the Pareto quasiordering says that x is equally morally good as y iff life-history (a_1, b_1) is equally good as life-history (a_1^*, b_1^*) and life-history (a_2, b_2) is equally good as life-history (a_2^*, b_2^*) and … and life-history (a_N, b_N) is equally good as life-history (a_N^*, b_N^*). It says that x is morally better than y iff life-history (a_1, b_1) is at least as good as life-history (a_1^*, b_1^*), life-history (a_2, b_2) is at least as good as life-history (a_2^*, b_2^*), and so forth, with at least one case in which (a_i, b_i) is strictly better than (a_i^*, b_i^*). Otherwise x and y are morally incomparable.

To see how various SWFs can extend the Pareto quasiordering, without necessarily producing a complete ordering, consider a case in which there are two individuals. Imagine, first, that $x = ((3,3), (3,3))$ while $y = ((4,4), (3,2))$. Then, according to $u(.)$ and all positive multiples, the utility vector corresponding to x is $(6r, 6r)$ and the utility vector corresponding to y is $(8r, 5r)$. According to $v(.)$ and all positive multiples, the utility vector corresponding to x is $(9s, 9s)$ while the utility vector corresponding to y is $(16s, 6s)$. Note that, in this case, the outcomes are Pareto noncomparable. However the utilitarian SWF, both using $u(.)$ and all positive multiples, *and* using $v(.)$ and all positive multiples, says that y is a better outcome. ($8r + 5r > 6r + 6r$ *and* $16s + 6s > 9s + 9s$).

On the other hand, if $x = ((3,3), (3,3))$ and $y = ((9,2), (1/3, 1/3))$, then the outcomes are both Pareto noncomparable and noncomparable according to the utilitarian SWF. Note that, according to $u(.)$ and all positive multiples, the utility vector corresponding to $x = (6r, 6r)$ while the utility vector corresponding to $y = (11r, 2/3r)$. According to $v(.)$ and all positive multiples, the utility vector corresponding to $x=(9s, 9s)$ while the utility vector corresponding to $y = (18s, 1/9s)$. Thus the utilitarian SWF, using $u(.)$ and all positive multiples, says that x is a better outcome (because $6r + 6r > 11r + 2/3r$). However, using $v(.)$ and all positive multiples, it says that y is a better outcome (because $9s + 9s < 18s + 1/9s$.)

Similar examples can be constructed for the other SWFs mentioned here (leximin, continuous prioritarian, rank-weighted).

I will reiterate that Chapter 3 will have to take on the burden of showing that well-being is indeed interpersonally comparable to some extent. I have tried to clarify the concept here, but have not yet shouldered that large burden. Even apart from the question of interpersonal comparisons, Chapter 3 will need to take on a large philosophical task: what *is* the most attractive account of well-being? Why believe (as I will ultimately argue) that well-being is a matter of idealized extended preferences? Moreover, how exactly will we arrive at an account of well-being representable by a set **U**?

Chapter 4 will then have to take on the burden of sorting *between* different SWFs. What is the most attractive rule R? The continuous prioritarian SWF, again, says: x is morally at

[42] In accordance with the account of well-being presented in Chapter 3, this could happen if some spectators have extended preferences over life-history lotteries and comparisons to nonexistence represented by $u(.)$ and positive multiples, while others have preferences represented by $v(.)$ and positive multiples.

least as good as *y* iff, for all *u*(.) belonging to **U**, $\sum_{i=1}^{N} g(u_i(x)) \geq \sum_{i=1}^{N} g(u_i(y))$, with *g*(.) strictly increasing and strictly concave. That means that *g*(.) takes the following shape:

[chart]

The continuous prioritarian SWF is not only anonymous and Paretian, but has a certain kind of "fairness" property, namely that it gives greater weight to well-being changes affecting worse-off individuals. This can be encapsulated in the *Pigou-Dalton* principle in terms of well-being, which says.

The Pigou-Dalton principle in terms of well-being

> Transferring a unit of well-being from a person at a higher well-being level to a person at a lower well-being level, without loss, and without the individuals switching ranks, improves the moral goodness of an outcome. In other words, *R* must be such that if we take a utility vector ($u_1, \ldots u_i, \ldots, u_j, \ldots u_N$), and alter it to ($u_1, \ldots, u_i + \Delta u \ldots, u_j - \Delta u, \ldots u_N$), where $\Delta u > 0$, $u_j > u_i$, and $u_j - \Delta u \geq u_i + \Delta u$, then *R* ranks the second vector above the first.

Moreover, as we'll see in Chapter 4, the continuous prioritarian SWF has a separability property and a continuity property. By contrast, the utilitarian SWF (which is unconcerned about fair distribution) lacks the Pigou-Dalton property. The rank-weighted SWF satisfies the Pigou-Dalton property but lacks the separability property. And the leximin SWF satisfies both of these properties, but gives absolute priority to worse-off individuals over better-off individuals. It refuses to make any tradeoffs – which means that it lacks the continuity property.

Sorting between these various SWFs, and further possibilities, will require much intellectual labor. All I have tried to do here is introduce the framework, so as to see its promise and so as to better understand the features of competing approaches.

<u>The Social Welfare Function Tradition in Welfare Economics and Contemporary Policy Analysis</u>

I have provided a definition of "the SWF framework," which sees it as one particular specification of the more generic architecture of a welfarist choice-evaluation procedure. Again, this definition is my own; the existing literature does not conceptualize the social welfare function in precisely this way, and, to keep things clear, I will use the abbreviation "SWF" only to refer to my particular definition. Still, my definition has many connections to the literature,

The notion of a social welfare function was introduced into economics by Abram Bergson in a 1938 article and was adopted by Paul Samuelson in Chapter 7 of his influential work, *Foundations of Economic Analysis*, published in 1947. My understanding of the SWF framework as way to outfit a *moral* decision procedure is in accordance with Bergson's and

Samuelson's views. Although these authors do not use the term "moral," they clearly do see the social welfare function as a way to encapsulate a set of *normative* judgments – and not just any old normative judgments, but the particular species of normative judgment I am referring to with the term "moral." Bergson writes that the social welfare function is a "scale of values for the alternative uses of resources," and also describes it as resting on "ethical" premises; I take "ethical," here, to be a synonym for "moral." Samuelson writes: "[W]e take as a starting point for our discussion a function [the social welfare function] of all the magnitudes of a system which is supposed to characterize some ethical belief – that of a benevolent despot, or a complete egotist, or 'all men of good will,' a misanthrope, the state, race, or group mind, God, etc."

Bergson is particularly clear in emphasizing that specifying the form of the social welfare function involves normative reasoning, and therefore has the characteristic features of such reasoning. He writes: "A notable feature in welfare economics is the attempt to formulate a criterion of social welfare without recourse to controversial ethical premises. …. [T]his goal for the criterion is an illusion." Or to put the point in terms of contemporary moral epistemology: there are a wide variety of logically possible SWFs, and choosing between them cannot be a matter of formal deduction, but should involve seeking a reflective equilibrium.

Bergson also states quite explicitly that the social welfare function is a tool that can be used to provide ethical advice to *anyone*, regarding "large" decisions, not merely to governmental decisionmakers.

> [I conceive] the problem of the [social welfare criterion] as having two aspects. First, it is necessary to determine the ethical values which one would take as data in counseling one or another citizen in any particular community on decisions involving alternative social states. Reference is not to decisions of the restricted sort taken in the market, but to decisions of the large sort usually implemented by actions of government, e..g, a tax reform. ….
>
> …
>
> I have been assuming that the concern of welfare economics is to counsel individual citizens generally. If a public official is counseled, it is on the same basis as any other citizen. In every instance reference is made to some ethical values which are appropriate for the counseling of the individual in question.

There are certainly subsequent authors who have adopted a different view of the social welfare function– for example, as providing a special kind of second-order normative advice that pertains only to governmental choice; or even as a descriptive rather than normative tool. Here, I should stress that the concept of the social welfare function can be employed to serve the particular aims of this book – helping to structure a decision procedure that can be used to *morally* evaluate anyone's large-scale choice – even if others in the social-welfare-function tradition have understood it as serving a different role. But it *is* heartening to note that my understanding seems close to that of the first movers in this tradition, Bergson and Samuelson.

Bergson and Samuelson stress that the social welfare function, in its most generic form, is defined directly on (roughly) what I term an outcome," namely a description of reality incorporating whatever features are taken to be of moral relevance.

> In dealing with this whole question [of value judgment], the present writer has found it useful to introduce into the analysis a welfare function, W, the value of which is understood to depend on all the variables that might be considered as affecting [the community's] welfare: the amounts of each and every kind of good consumed by and service performed by each and every household, the amount of each and every kind of capital investment undertaken, and so on. The welfare function is understood initially to be entirely general in character; its shape is determined by the specific decisions on ends that are introduced into the analysis. Given the decisions on ends, the welfare function is transformed into a scale of values for the evaluation of alternative uses of resources.

Admittedly, Bergson and Samuelson, as well as subsequent scholars, have given too little attention to the point that a social welfare function meant to be used as a decision-making tool must be cognitively tractable, and thus that its arguments cannot be anything like *complete* descriptions of possible realities. Thus I have stressed that an "outcome" is a *simplified* possible world.

In any event, modulo this point about simplification, the Bergson-Samuelson social welfare function in its most generic form is defined directly on outcomes. It takes the form $w(x)$; the numerical fact $w(x) > w(y)$ embodies the moral ("ethical") judgment that x is a morally ("ethically") better outcome than y. Further, it provides a *complete* quasiordering, i.e., a complete, reflexive and transitive ordering of the outcomes. As Samuelson explains, the "ethical" belief which the social welfare function reflects must be "such as to admit of an unequivocal answer as to whether one configuration of the economic system is 'better' or 'worse' than any other or 'indifferent,' and … these relationships [must be] transitive; i.e., *A* better than *B*,*B* better than *C*, implies *A* better than *C*, etc." I too will see the SWF as furnishing a quasiordering of outcomes – but not necessarily a complete one.

Bergson and Samuelson then explain that this generic form $w(x)$, coupled with the further "ethical" judgment that the goodness of outcomes depends on individuals' preferences, yields a social welfare function whose argument are individual *utilities* representing each individual's preference-satisfaction. It becomes $w(u_1(x), u_2(x) , \ldots , u_N(x))$.

Bergson and Samuelson believed that ranking outcomes using a social welfare function with individual utilities as its arguments did not presuppose the possibility of interpersonal comparisons. Indeed, they believed that a social welfare function could not only "extend" the Pareto quasiordering, but could indeed yield a complete ordering of outcomes, even if well-being was not interpersonally comparable.

However, this view was seriously called into question by Kenneth Arrow's earth-shattering work, *Social Choice and Individual Values*. Arrow addressed himself directly to the problem of formulating a social welfare function. For Arrow, as for Bergson and Samuelson,

this device will produce a complete quasiordering of what he calls "states," i.e., "outcomes" (modulo simplification):

> The most precise definition of a social state would be a complete description of the amount of each type of commodity in the hands of each individual, the amount of labor to be supplied by each individual, the amount of each productive resource invested in each type of productive activity, and the amounts of various types of collective activity, such as municipal services, diplomacy and its continuation by other means, and the erection of statues to famous men.

But, importantly, Arrow formulates the problem differently than Bergson and Samuelson. What Arrow calls a "social welfare function" is not a mathematical function of the form $w(u_i(x), u_2(x), \ldots, u_N(x))$. It takes as its input, not individual utility numbers, but, rather, each individual's preference ranking. The Arrow set up is this: Each individual's preferences take the form of a complete quasiordering of the set of outcomes; and the "social welfare function" is a rule that maps this profile of individual orderings (one for each of the N individuals in the population) onto a social ordering .

> By a social welfare function will be meant a process or rule which, for each set of individual orderings $R_1, .. R_n$ for alternative social states (one ordering for each individual), states a corresponding social ordering of alternative social states, R.

Arrow proved that it was impossible to produce such a "process or rule" that satisfied several seemingly innocuous axioms, namely nondictatorship, universal domain, a Pareto condition, and the "independence of irrelevant alternatives."

Since Arrow's book, theoretical work regarding social welfare functions has proceeded in two directions. One body of scholarship retains the basic Arrow set-up. It seeks to produce a social ranking of outcomes as a function of the profile of individual preference rankings – investigating the possibility of doing so if one or more of the Arrow conditions are relaxed, in particular universal domain and the independence of irrelevant alternatives. In substantial part, the conclusions of this literature have been to confirm Arrow's impossibility result even with such relaxation. However, the literature is not completely negative. One positive result is the so-called "egalitarian equivalent" ordering. This rule for ordering outcomes assumes no interpersonal comparability, is sensitive to fair distribution, and gets around the Arrow result by relaxing the "independence of irrelevant alternatives " condition. It will be discussed later in the chapter, in connection with inequality metrics.

A second body of theoretical scholarship commenced in the early 1970s, triggered by Amartya Sen's 1970 work *Collective Choice and Social Welfare*. This literature returns to the original Bergson/Samuelson idea of making the social ordering of outcomes a function of individual *utilities*. Unlike Bergson and Samuelson, however, this literature is not resistant to the notion of interpersonal welfare comparisons.

More precisely, the set-up in this second body of post-Arrovian work is as follows. There is a set of social states (what I am calling an outcome set **O**) and a set of *N* individuals. Each individual *i* has a *utility function* $u_i(.)$, where each such $u_i(.)$ maps the various states onto real numbers. Given the *N* individuals, there is a profile of individual utility functions $[u_1(.), \ldots u_N(.)]$. The central problem that this literature focuses upon is arriving at a complete quasiordering of the social states as a function of this profile of individual utility functions.

By moving from the Arrow set-up (a profile of individual preference orderings) to the newer set-up (a profile of individual utility functions) , this literature accommodates the possibility of interpersonal comparisons.[43]

> Central to Arrow's impossibility theorem is the inability of a social welfare function to use any information beyond that given by the individual preference orderings on the set of possible alternatives. In particular, the very formulation of the problem rules out the use of *interpersonal comparisons* of well-being or utility. Classical social decision rules such as utilitarianism, or any other rule which allows for tradeoffs between the utilities experienced by different individuals, simply cannot be expressed in terms of an Arrovian social welfare function….
>
> In order for the social choice procedure to incorporate information about interpersonal comparisons of utility, the notion of a social welfare function has to be generalized. Instead of determining the social preference on the basis of a profile of individual preference orderings, a *social welfare functional* assigns a social preference to each admissible profile of individual *utility functions*. Different assumptions concerning the measurability and interpersonal comparability of utility can be formalized by partitioning the set of admissible profiles of utility functions into sets of informationally equivalent profiles and requiring the social welfare functional to be constant on a cell of the partition.

The SWF framework, as I have defined it, is very closely related to this literature. That framework yields a quasiordering of outcomes by using a set **U** of utility functions and a rule *R* for ranking utility vectors , applied as follows: *x* is at least as good as y iff $u(x)$ is ranked at least as good as $u(y)$ by *R*, for all $u(.)$ belonging to **U**. Each $u(.)$, in my framework, is a vector-valued function that maps a given outcome onto an *N*-entry list of numbers. The literature now under discussion uses the idea of a *profile* of individual utility functions $[u_1(.), \ldots u_N(.)]$, each of which maps a given outcome onto a single real number. So a particular utility function, in my setup, is just the same thing as what this literature sees as a particular profile of individual utility functions.

Moreover, as the last sentence of the immediately quoted paragraph suggests, a central focus of this literature is the problem of *invariance*. An invariance requirement says that, if combining an SWF with a given profile of individual utility functions $[u_1(.), u_2(.), \ldots u_N(.)]$

[43] It might be asked whether scholars working in this particular literature see individual utilities as representing individual preference satisfaction and, thus, individual well-being (on the assumption that well-being reduces to preference satisfaction) or as representing preference satisfaction independent of well-being (e.g., the satisfaction of moral preferences). See supra note ___. It seems that many in this literature take the first view (as suggested by the quotation here in the text). In any event, whatever the interpretation of the social welfare function that some in the literature may adopt, it provides a formal structure that I can use for my welfarist purposes.

produces a particular ordering of outcomes, then the ordering produced by combining that SWF with a stipulated kind of transformation of that profile, $[\varphi_1(u_1(.)),\varphi_2(u_2(.)),\ldots,\varphi_N(u_N(.)]$, must be the same. For example, we might say that the ordering of outcomes must not change if a given profile is transformed by a "positive affine transformation" (multiplying each function in the profile by a positive constant and adding a constant). Or we might say that the ordering of outcomes must not change if a given profile is transformed by applying a so-called "increasing transformation" to each function in the profile.

In effect, then, this literature thinks of the ordering of outcomes as being produced by the combination of an SWF and a *set* of profiles: a given profile, plus all the profiles that are equivalent given some invariance requirement. My approach, too, sees the ordering of outcomes as being generated by a combination of an SWF and a *set* of vector-valued utility functions, i.e., a set of profiles.

There *are* some differences between my approach and this literature's. In particular, the literature focuses on the possibility of deriving a *complete* ranking of outcomes from a set of profiles of individual utility functions. In other words, this literature might be understood as adopting the SWF framework as I have defined it, with the additional stipulation that: R and **U** must be such that the rule, x is at least as good as y iff $u(x)$ is ranked at least as good as $u(y)$ by R, for all $u(.)$ belonging to **U**, yields a complete quasiordering. For reasons I have elaborated, I see no reason to insist on completeness.

Further, the literature sees the utility functions as representing preferences – the number $u_i(x)$ represents something about individual i's preference-satisfaction in outcome x -- while I see the utility functions in **U** as more generic. They represent *some* account of well-being, be it preferentialist, mental state, or objective.

My account of interpersonal comparisons is closely related to the literature's, but also more general. Within my framework, whether an account of well-being allows for interpersonal or intrapersonal comparisons will "show up" in **U**. And that is, in effect, what the literature says as well.[44] But I allow for different *degrees* of level or difference comparability. Some life-

[44] Strictly, the literature associates different kinds of comparability with difference invariance requirements. For example, it says that interpersonal level comparability means that the SWF is invariant to a common increasing transformation of all individual utility functions. Interpersonal level and difference comparability means that the SWF is invariant to a common positive affine transformation of all individual utility functions.

But to say the first is just to say that the SWF should produce a complete ordering of outcomes if **U** consists of a $u(.)$ and all increasing transformations thereof. If we employ the natural rule for using **U** to represent the ordering of life-histories (that a life-history is at least as good as a second iff all $u(.)$ rank it at least as high), then a **U** consisting of a $u(.)$ and all increasing transformations thereof will indeed render comparable all life-histories, including life histories of different individuals. Level comparability "shows up" in the **U** generated by the invariance requirement. Similarly, if **U** consist of a $u(.)$ and all positive affine transformations thereof, and we employ the natural rule mentioned earlier for using **U** to represent differences, then this **U** will render comparable all life-histories and all differences between life-histories, including life histories of different subjects. Now, level and difference comparability "show up" in the **U** generated by the invariance requirement.

histories might be comparable, while others aren't. By contrast, the literature tends to assume that level or difference comparability is "all or nothing."

Finally, the literature doesn't really grapple with the problem of how to make interpersonal comparisons. Rather, it focuses on the important mathematical problem of identifying the possible forms of the rule *R*, given a set **U** which allows for a certain degree of comparability, and given different possible stipulations about the ordering of outcomes. To give one example, it might ask: given that **U** consists of a single utility function and all positive affine transformations thereof, if we require that the SWF be such as to produce a ranking of outcomes which is complete and is separable across persons, what are the possible forms of *R*? The more philosophical question concerning how we arrive at that particular set **U**, from an underlying account of well-being (preference based or otherwise) is not confronted.

Still, the SWF framework as I present it is a really just an incremental extension of what this literature understands as a social welfare function. Moreover, the literature's axiomatic results about the possible forms of *R* will come into play at many junctures in the analysis.

The body of scholarship that I have been discussing is purely theoretical. But the concept of a social welfare function also plays a role in a number of other literatures, which use that concept – at least in part -- to formulate policy guidance. Perhaps the most important example is the field of "optimal tax theory," spurred by James Mirrlees' 1971 article, "An Exploration in the Theory of Optimum Income Taxation," for which he ultimately won the Nobel Prize. In this article, Mirrlees addressed the very difficult problem of setting an income tax schedule, given the heterogeneity of individuals' abilities, the incomplete observability of individual abilities to the tax authorities, and the disincentive effects of income taxes on labor. He did so in a much more rigorous fashion than had previously been accomplished by seeking to identify the tax schedule that maximized a social welfare function, taking as its arguments the array of individual utilities associated with each outcome; each individual's utility, in turn, was a function of that individual's consumption and labor.

"Optimal tax scholars" pursue the approach that Mirrlees pioneered. In this field, now quite substantial, economists investigate a particular policy domain (taxes), via social welfare functions defined on individual utilities; and they do so on the assumption that individual well-being is interpersonally comparable. Mirrlees characterizes the social welfare function as "embodying interpersonal comparisons of welfare." Matti Tuomala, surveying the field of optimal tax theory, observes that it is "firmly based on explicit interpersonal comparisons" It also bears note that the main functional form for analyzing tax policies, in the optimal tax literature, is a continuous prioritarian form – and indeed, more specifically, the Atkinsonian social welfare function I will argue for in detail in Chapter 4.

Social welfare functions have also been used or at least discussed as a policy-analysis tool in a number of other scholarly literatures, including: optimal growth theory; environmental

economics[45] (including several major recent studies on climate change); the estimation of cost-of-living indices; health economics (a number of scholars have proposed using social welfare functions to evaluate health policy, with utility numbers representing individuals' health states); and legal scholarship (Kaplow and Shavell argue in a prominent book, *Fairness versus Welfare*, that legal scholars should use SWFs to evaluate legal doctrines). Further, an important strand of the inequality measurement literature sees a close link between social welfare functions and inequality measures. As we'll see, the Atkinson *inequality metric* is derived from the Atkinson *SWF*. Similarly, there turns out to be a connection between the Gini coefficient (another standard inequality metric) and the rank-weighted SWF (a different kind of SWF).

Finally, although social welfare functions are not (as far as I'm aware) currently employed by governmental bodies, there are a few governments that engage in cost-benefit analysis with *distributive weights* – an approach which is a variation on the social welfare function approach.

In short, the idea of a social welfare function has a rich background in economic theory, and also has gone well beyond the stage of theoretical elaboration. Still, the approach is not nearly as widely used in contemporary policy analysis as the competing approaches, to which we now turn.

Cost-Benefit Analysis

Cost-benefit analysis ("CBA") has come to dominate policy analysis within the U.S. federal government. Since 1981, federal administrative agencies subject to Presidential control have been required to prepare a full CBA whenever a sufficiently large-scale regulation is issued. These documents – which are typically voluminous, and can cost millions of dollars to prepare – are reviewed by a powerful oversight body within the Office of Management and Budget. Similar systems of "regulatory review" have existed for some time in Australia and, more recently, have been put in place in Britain, the EU, and other jurisdictions. CBA also has an important role in many governments outside the regulatory-review context. For many years, dating from the mid-1930s, CBA has been the key methodology employed by the U.S. federal government to evaluate dams and flood-control measures. More generally, it is widely used both in the U.S. and abroad to evaluate large infrastructure projects.

Within the scholarly field of economics, CBA has an odd status. On the one hand, it is generally viewed critically by theoretical welfare economists, who are keenly aware of the various difficulties with the approach and the closely related Kaldor-Hicks tests -- difficulties that I shall discuss below. Yet cost-benefit analysis is the foundation for modern *applied* welfare economics. Most contemporary economists who actually evaluate governmental policies, or undertake related empirical work, do so using CBA as the basic evaluation tool. For example, vast bodies of scholarship by applied economists are devoted to estimating

[45]Once more, here, the Atkinsonian SWF is central.

willingness-to-pay/accept (WTP/WTA) amounts, the inputs for CBA, either by inferring WTP/WTA amounts from demand curves and other "revealed preference" data, or by conducting survey research. Within the broader policy-analysis community – including not academic economists, but also researchers at "think tanks" and the like – CBA is similarly widespread.

What explains the dominance of CBA within applied economics, despite the widespread view among theoretical welfare economists that it lacks convincing foundations? This is an interesting sociological and historical question. It seems that the answer has much to do with the unease that many economists feel about making interpersonal welfare comparisons. Kaldor-Hicks tests were constructed at a time (the 1930s and 1940s) when such comparisons were universally disfavored within economics. It soon became clear that those tests were problematic, and a subliterature within theoretical economics eventually emerged, a generation later, in which interpersonal comparisons were accepted and interpersonally comparable utilities used as inputs to social welfare functions. Around the same time, "optimal tax" scholars began to utilize social welfare functions in their work; and social welfare functions are now also used in certain other areas of economics. There *are* therefore a number of contemporary economic literatures where interpersonal comparability is embraced. But the suspicion of interpersonal comparisons has remained widespread in the profession at large – so much so that many economists view welfare economics as being a "dead" field. In point of fact, welfare economics is not "dead" at all; interesting and important theorizing continues apace. What the characterization indicates, instead, is that many economists outside the fields where social welfare functions are accepted -- many applied welfare economists, and many economists whose work is positive rather than normative – disagree with the proponents of that approach on the very basic question of interpersonal welfare comparability.

So much for background. Let us now focus on CBA itself. As already stated, the key building block for CBA is the construct of a WTP/WTA amount.[46] This construct presupposes that outcomes are characterized, at least, in terms of individuals' consumption. More specifically, let us imagine that outcomes are described as having a single period, characterized in terms of each individual's consumption plus perhaps other individual attributes and background facts. For short, each outcome has the form $(c_1,\ldots, c_N, \mathbf{b})$.[47] The **b** term, here, encompasses all individual non-consumption attributes and background facts that are included in the description of the outcome (such as individuals' health states, the levels of public goods, environmental quality, individuals' happiness states, the price vector, etc.). Taking outcome *x* as baseline, individual *i*'s WTP/WTA for *y* is the change in his consumption in *y* which would make him indifferent between *x* and *y*. In other words, if $x = (c_1,\ldots, c_N, \mathbf{b})$ and $y = (c_1^*,\ldots, c_N^*,$

[46] This is sometimes referred to as an equivalent or compensating variation.

[47] As already explained, the generic form of the characterization of some period in an outcome is $(\mathbf{a}_1^t, \mathbf{a}_2^t, \ldots, \mathbf{a}_N^t, \mathbf{a}_{imp}^t)$, where $\mathbf{a}_i^t$ is a vector of attributes of individual *i* during period *t*, and $\mathbf{a}_{imp}^t$ are background facts about the world.

b*), then individual *i*'s WTP/WTA for outcome *y*, taking *x* as baseline, is the amount Δc_i such that *i* is indifferent between $=(c_1^*,\ldots,c_i^* - \Delta c_i,\ldots, c_N^*, \mathbf{b}^*)$ and $(c_1,\ldots, c_i, \ldots, c_N, \mathbf{b})$.

Welfarist decisionmaking can also, of course, employ outcomes that have multiple periods. CBA readily generalizes to the multi-period setup – as do the criticisms I will present – but to simplify the presentation I will focus on the one-period case. I similarly simplify the presentation, later in the chapter, in discussing other policy-analysis approaches. Thus, throughout the chapter, I will talk about an individual having a single level of some attribute (a single consumption amount, health amount, leisure level, etc.) in an outcome, or background facts being set at a particular level. This is meaningful in the case of a single-period outcome but not in the case of a multi-period outcome -- where we would need to distinguish between the individual's consumption during one period and her consumption during another period, and so forth for other attributes or background facts.

Back to the one-period setup and the definition of CBA in that set-up. Here, an individual's consumption in some outcome means the total value, in terms of market prices in that outcome, of the array of marketed goods and services that she uses up. If the physical quantities of goods or services consumed by individual *i* in outcome *x* are $q_i^1,\ldots,q_i^M$, and the prices of those goods or services are $p^1,\ldots,p^M$, then individual *i*'s consumption in *x* is $p^1 q_i^1 + \ldots + p^M q_i^M$. (In the highly simplified case where individuals are modeled as consuming but a single good, q^1, individual *i*'s consumption is simply q_i^1). Another term that economists often use for consumption is "expenditure." Note also that in the one-period setup, at least absent bequests, there is no difference between an individual's income and his consumption. The terms "income," "expenditure," and "consumption" are therefore often used as synonyms -- which is how I will use those terms throughout the book, unless otherwise noted. [48]

[48] The reader might wonder how WTP/WTA amounts can be well-defined, once we move away from the super-simple context of a single consumption good. Imagine that $x = (c_1,\ldots,c_N, \mathbf{p}, \mathbf{b})$ and that $y = (c_1^*,\ldots,c_N^*, \mathbf{p}^*, \mathbf{b}^*)$. There are *M* different marketed goods. In outcome *x*, the vector of prices of those *M* goods is **p**; in outcome *y*, the vector of prices of those *M* goods is **p***. In this case, c_i is the total value of the goods that *i* consumes in *x*, as calculated using the vector of prices **p**; c_i^* is the total value of the goods that *i* consumes in *y*, as calculated using the vector of prices **p***. How can we determine that individual *i*'s WTP/WTA amount is a particular amount Δc_i? To know whether *i* is different between $(c_1,\ldots, c_i, \ldots c_N, \mathbf{p}, \mathbf{b})$ and $(c_1^*,\ldots c_i^* - \Delta c_i,\ldots,c_N^*, \mathbf{p}^*, \mathbf{b}^*)$, don't we need to know what particular array of goods *i purchases* in the two outcomes?

The answer is that CBA builds in a particular behavioral assumption. The assumption is that, in all outcomes, individuals are preference-maximizers. For a given price vector, and a given amount of consumption – i.e., total expenditure on marketed goods – CBA assumes that this expenditure is allocated between the goods so as to maximize the individual's preference-satisfaction. Formally, this is handled by assigning each individual a "direct" intrapersonal utility function defined on vectors of consumption goods; by deriving from that an "indirect" utility function with income and the price vector as its arguments, i.e., the maximum achievable direct utility for a given income and price vector; and calculating WTP/WTA amounts using indirect utility.

WTP/WTA amounts are defined – it bears emphasis – in terms of individual's preferences, which are assumed to be complete. For each pair of outcomes, it is assumed, each individual either weakly prefers the first to the second, or weakly prefers the second to the first, or both; this relation of "weak preference" is complete, reflexive, and transitive, and gives rise to relations of indifference and strict preference that are, respectively, reflexive, symmetric, and transitive and irreflexive, asymmetric, and transitive. It also bears emphasis that WTP/WTA amounts are traditionally understood in terms of individuals' actual (non-ideal) preferences. Although individuals *are* assumed to satisfy basic rationality properties, there is no assumption that the preferences used to determine WTP/WTA amounts are idealized in any strong sense.

We have discussed how to calculate WTP/WTA amounts for each individual. The basic CBA test is then simply to add these. Given two outcomes, x and y, pick one as baseline (x); and say outcome y is better than x if the sum of WTP/WTA amounts is positive, worse if the sum is negative, and equally good if the sum is zero.

CBA has wide potential applicability – as evidenced by the very wide array of policy areas where it is employed, by analysts inside and outside government. CBA *measures* each individual's well-being change in terms of a consumption change, but the technique in no way requires that outcomes be characterized solely in terms of consumption. CBA can, in principle, be used to rank outcomes that include many different sorts of non-market characteristics. Or, at least, CBA can be so used if preferences are "well-behaved" – a point to be discussed in a moment. CBA also, clearly, does not presuppose the possibility of interpersonal welfare comparisons. A given individual's WTP/WTA amounts for various outcomes, relative to some baseline outcome, are a function of her preferences over outcomes – in other words, of her ranking of her own life-histories. We can determine such amounts regardless of whether life-histories involving different subjects can be ranked in terms of well-being.

However, I shall argue that CBA is not an attractive welfarist procedure for morally evaluating governmental policies and other large-scale choices – except in the case of distributively weighted CBA, which adjusts WTP/WTA amounts using weighting factors and is just a variation on the SWF approach. I will ignore distributively weighted CBA for the moment, and focus on the standard approach to CBA, which aggregates WTP/WTA amounts without distributive weights. My discussion (as throughout this chapter) concerns difficulties in

---

However, the specific behavioral assumption here – that individuals act in each outcome to maximize their preferences -- is actually not essential to CBA. All that is needed is *some* set of behavioral regularities that determine, in each outcome, which array of marketed goods and services individual i will purchase – as a function of the individual's consumption c_i, the price vector $\mathbf{p}$, and perhaps other characteristics of the individual or outcome. With those behavioral regularities in hand, we can map variations in individual i's consumption in a given outcome y onto variations in the array of marketed goods and services that he purchases, and thus determine individual i's WTP/WTA for y relative to baseline x (Δc_i).

producing an acceptable moral ranking of an outcome set. I discuss, first, the difficulties in using CBA to construct an outcome ranking that satisfies minimal welfarist criteria of being a quasiordering that respects Pareto indifference and Pareto superiority. I then argue that the standard Kaldor-Hicks defense of CBA is very unpersuasive. I next discuss a possible revisionary defense of CBA that accepts the possibility of interpersonal comparisons and sees CBA as a proxy for overall well-being. Finally, I describe some plausible roles that CBA might play, other than as a criterion of the moral goodness of outcomes, and clarify that nothing in the discussion here is meant to undercut those possible roles.

Minimal Welfarist Criteria

As explained, traditional CBA defines WTP/WTA amounts in terms of actual preferences, and estimates policy impacts by estimating the sum of WTP/WTA amounts, thus defined. But what if our welfarist decisional framework is built around some alternative account of well-being *W*, which we take to be more attractive than an actual-preference-based theory? If so, CBA can clearly violate the principles of Pareto indifference or Pareto superiority, understood in terms of *W*.

For example, imagine that *W* analyzes individual well-being in terms of idealized (fully-informed, fully-rational) extended preferences, along the lines I will defend in the next chapter. In that case, two outcomes *x* and *y* are Pareto indifferent (in terms of *W*) if, for every individual *k*, and every pair of life-histories (*x*; *i*) and (*y*; *i*), individual *k* under idealized conditions would be indifferent between the two life-histories.[49] But it is quite possible for *x* and *y* to be Pareto-indifferent in this sense and yet for some individuals to actually prefer *y* over *x*, or vice versa -- in which case those individuals will have positive or negative WTP/WTA amount for *y*, and thus the sum of WTP/WTA amounts for *y* could be nonzero, producing a ranking of *x* and *y* that violates Pareto indifference. Individual *i*'s *actual* preferences, regarding *x* and *y*, can readily deviate from her *idealized, self-interested* preferences regarding those outcomes. Individual *i* might be poorly informed, less than fully rational, or non-self-interested.

Similarly, imagine that *W* analyzes individual well-being in terms of some list of objective goods. In that case, two outcomes *x* and *y* are Pareto indifferent (in terms of *W*) if each individual is just as well off in the two outcomes, in terms of the goods. But even if *x* and *y* are Pareto indifferent in this sense, an individual's actual preferences can deviate from the list of goods, and individuals can have positive or negative WTP/WTA amounts for *y*. Finally, imagine that *W* analyzes individual well-being in terms of good or bad mental states. In that case, *however* exactly the goodness and badness of mental states is specified, outcomes *x* and *y* must be Pareto-indifferent, in terms of *W*, if each person's mental states in *x* are the same as her

[49] More precisely, as explained in the next chapter, this means that *i* herself, if fully rational, fully informed, and self-interested, must be indifferent between outcomes *x* and *y*; and that other individuals, if fully rational, fully informed, and focused on *i*'s well-being, must be indifferent between the outcomes.

mental states in *y*. But individuals can actually prefer features of outcomes other than their own mental states, and thus the sum of WTP/WTA amounts for *y* can be nonzero.

I have been discussing how the sum of WTP/WTA amounts can yield violations of *Pareto indifference* understood in terms of accounts of well-being other than an actual-preference account. A parallel analysis will show how the sum of WTP/WTA amounts can yield violations of *Pareto superiority* understood in terms of accounts of well-being other than actual-preference accounts.

Of course, the proponent of CBA might respond that an actual-preference-based account of well-being simply *is* the best account of well-being. But there is very good reason to reject this response: strong critical arguments can be mounted (and will be mounted in Chapter 3) against an actual-preference-based theory of well-being. Even if the reader rejects my affirmative claim that the most attractive *W* is an ideal-preference account, there remains very good reason for concluding that the most attractive *W* is not an actual-preference account; and thus that the sum of WTP/WTA amounts can run afoul of Pareto indifference and superiority in terms of *W*, whatever exactly it may be.

A solution, here, is to construct a more generic version of CBA. For any given outcome set **O**, and any account of well-being *W*, we can define WTP/WTA amounts in terms of *W*. Let us say that individual *i*'s WTP/WTA amount for outcome *y*, relative to outcome *x*, is the change in individual *i*'s consumption in *y* that makes him equally well off as in *x*, as judged by *W*. Formally, if $x = (c_1,\ldots, c_N, \mathbf{b})$, and $y = (c_1^*,\ldots, c_N^*, \mathbf{b^*})$, then individual *i*'s WTP/WTA amount for *y* is the amount Δc_i such that life-history $(c_1,\ldots, c_i, \ldots, c_N, \mathbf{b}; i)$ is ranked by *W* as equally good as life-history $(c_1^*,\ldots,c_i^* - \Delta c_i, \ldots, c_N^*, \mathbf{b^*}; i)$.

This more generic definition of WTP/WTA amounts will at least give us a shot at using CBA as a basis for producing a Pareto-respecting quasiordering of outcome sets – regardless of what theory of well-being the welfarist takes to be most attractive. Generic CBA is *modular* : if we are using an actual-preference-based account of well-being, it reduces to traditional CBA, but in other cases it generates WTP/WTA amounts so as to track the theory of well-being on hand.

I should note that the idea of redefining WTP/WTA amounts in terms of an account of well-being other than an actual-preference-based theory is not merely a theoretical proposal. Within the burgeoning field of "happiness" studies, a number of researchers have defined WTP/WTA amounts in terms of *happiness* rather than preference-satisfaction, and have employed happiness surveys to estimate WTP/WTA amounts thus defined.

However, even the generic version of CBA encounters some challenges in meshing with a given theory of well-being *W* and an outcome set to produce a Pareto-respecting quasiordering of that outcome set. One issue concerns the completeness of the ordering of individual life-histories. Traditional CBA is built on the premise that each individual has a complete preference ranking of her own life-histories; but the most attractive *W* may well *not* provide a complete

well-being ranking of each individual's life-histories. There may be pairs of life histories, (*x*; *i*) and (*y*; *i*), both involving a given individual *i*, such that two are incomparable according to *W*. If so, defining a single WTP/WTA amount for *y* relative to *x* poses a technical challenge (although one that may be overcome along lines elaborated in the margin).[50]

A second issue concerns lexical orderings of non-market goods relative to consumption. This is distinct from the problem of completeness just mentioned. Imagine that *W* does generate a complete ranking of each individual's life-histories. However, this ranking is such that there are certain reductions in the level of some non-market attribute that cannot be compensated for by increased consumption. If so, there will be cases in which individual *i* has no finite WTP/WTA amount for outcome *y* relative to *x*.[51]

Let us now bracket these issues and assume, instead, that the nexus between well-being and consumption is "well-behaved." By "well-behaved," I mean something like the following: (1) *Complete intrapersonal comparisons*. Our theory of well-being *W* is such that any life-history involving a given individual is ranked as better than, equally good as, or worse than any other life-history involving the same individual. (2) *Determinate WTP/WTA amounts that point in the same direction as well-being*. For any two outcomes *x* and *y*, if *i* is better off in *y* than *x*, then *i*'s WTP/WTA amount for *y* relative to *x* is a single, finite, positive amount. And if we substitute "worse off" or "equally well off" for "better off" in the preceding sentence, "positive" becomes "negative" or "zero." (3) *Well-being is strictly increasing in consumption.* For any triple of outcomes *x*, *y*, *z*, if *i* is better off in *z* than *y*, the WTP/WTA amount for *x* relative to *y* is greater than the WTP/WTA amount for *x* relative to *z*.[52]

Even with a well-behaved nexus between well-being and consumption, WTP/WTA amounts may be subject to *reversals*. By this I mean that *i*'s WTP/WTA for outcome *y*, relative to outcome *x*, may not be same magnitude (absolute value) as *i*'s WTP/WTA for outcome *x*,

---

[50] If *W*'s ranking of life-histories is representable by a set of utility functions **U**, then we can define a WTP/WTA amount, for a pair (*x*; *i*) and (*y*; *i*), for each utility function in the set **U**; we can assign each outcome a sum-of-WTP/WTA amount for each such utility function; and we can use some version of a CBA test that ranks outcomes depending on the set of aggregate benefit amounts thus assigned to outcomes.

[51] These two issues are not just theoretical niceties. CBA researchers have encountered many difficulties estimating WTP/WTA values for non-market goods, which may in part be a result of incomparabilities and lexical orderings in individuals' ranking of life histories.

Admittedly, with respect to lexicality, it is quite possible that this is being driven by individuals' moral preferences, and that a theory of well-being which washes out such preferences (as does the theory I will argue for in the next chapter) would not see non-market attributes as lexically ordered vis a vis consumption. Further, it must be conceded that lexicality in the well-being ordering of life-histories might well undermine the SWF approach -- by undermining the use of utilities to represent well-being. See infra Chapter 3, note ___. For these various reasons, lexicality of non-market goods vis a vis consumption should not be seen as a major element in the case for using SWFs rather than CBA.

[52] The last condition (which is not entailed by the first two) will be needed to ensure that the "money metric" style CBA test to be constructed in a few paragraphs yields a Pareto-respecting quasi-ordering.

relative to outcome *y*. This can readily arise in the case of an actual-preference-based theory of well-being; has been much discussed by the CBA literature; and would presumably generalize to alternative theories of well-being. WTP/WTA reversals can occur because the prices of marketed goods are different in *y* than *x*; or they can occur, even without price change, if individuals have different levels of non-market goods in the two outcomes.

Because of WTP/WTA reversals, it is possible that the *sum* of WTP/WTA values for outcome *y*, relative to outcome *x*, will be positive, and yet the sum of WTP/WTA values for outcome *x*, relative to outcome *y*, will also be positive. Thus, clearly, we cannot produce a quasiordering of an outcome set by a naïve use of the WTP/WTA criterion: by saying that *y* is morally better than *x* iff the sum of WTP/WTA amounts for *y*, relative to *x*, is positive. A "morally better than" relation, generated in this manner, will not necessarily be asymmetric. There will be cases in which *y* is morally better than *x*, but *x* is morally better than *y*. Yet, at a minimum, the "morally better than relation" should be asymmetric.

A more sophisticated application of the WTP/WTA test ranks a given outcome set by arbitrarily picking some baseline *x*; by calculating the sum of WTP/WTA amounts relative to that baseline; and by ranking the outcomes in the order of these total amounts. But this approach is not yet sophisticated enough. It has been shown that this approach can violate Pareto superiority, at least if WTP/WTA amounts are defined in the traditional way, in terms of actual preference satisfaction; and the problem would presumably generalize to various other accounts of well-being. It is possible that every individual has complete, well-behaved preferences; that *z* is Pareto superior to *y*; and yet that the sum of WTP/WTA amounts for *z*, taking *x* as baseline, is actually less than the sum of WTP/WTA amounts for *y*, taking *x* as baseline.

It might be argued that these "reversal"-based objections to CBA may not arise in many outcome sets. Reversals are theoretical possibilities, but in practice (given a well-behaved consumption/well-being nexus and generically defined WTP/WTA amounts in terms of whatever theory of well-being *W* is at hand) the WTP/WTA tests just mentioned will typically suffice to yield a Pareto-respecting quasiordering of an outcome set. Or so the argument might go.

The pervasiveness of reversals is a matter of some dispute. In any event, that problem can be circumvented by a yet more sophisticated application of the WTP/WTA test than the possibilities yet described. (The idea, here, is based on the notion of "money metric" utility, which some researchers have proposed as a way to solve some of the ordering anomalies associated with CBA.) For a given outcome set, and a given theory of well-being *W*, arbitrarily choose some outcome *x*. Then, for every other outcome *y*, determine each individual's WTP/WTA for *x*, taking *y* as baseline. If the nexus between well-being and consumption is well-behaved, given *W*, this amount will exist. Sum these amounts, across individuals, and assign *y* a net benefit value equaling the negative of the sum. Call this the "monetized aggregate benefit" of *y*. Order the outcomes in the order of their monetized aggregate benefits, with *x* itself assigned a monetized aggregate benefit of zero. It is not too hard to show that this ordering will

be a complete ordering; that it will assign the same monetized aggregate benefit number to Pareto-indifferent outcomes (so that it respects the principle of Pareto indifference); and that it will assign a larger monetized aggregate benefit number to *z* rather than *y* if z is Pareto superior to *y* (so that it respects the principle of Pareto-superiority).

To sum up: substantial challenges do arise in using CBA to construct a Pareto-respecting quasi-ordering of any given outcome set, given some theory of well-being *W*. But, with WTP/WTA defined in terms of *W*, there does seem to be a reliable way to meet this goal, via a sufficiently sophisticated application of the sum-of-WTP/WTA test – at least if the nexus between well-being and consumption is well-behaved.

Justifying CBA as a Principle for Ordering an Outcome Set: The Kaldor-Hicks Justification

Granted that we can use CBA to rank an outcome set in a manner that satisfies minimal welfarist criteria, what would *justify* this ranking? CBA will end up ranking the Pareto-noncomparable outcomes in an outcome set one way; a different welfarist approach will rank them a different way. Why believe that the ranking generated by CBA is the morally attractive ranking?

The traditional justification for CBA appeals to the construct of a *potential Pareto improvement*. This construct was developed during the 1930s and 40s, by leading figures in economics (in particular Kaldor and Hicks), at a point when these scholars were attempting to rebuild the field of welfare economics without using interpersonal welfare comparisons – a notion that had become strongly disfavored. The basic idea was to appeal to the possibility of transforming *y* into an outcome Pareto-superior to *x*, as a basis for judging *y* to be a better outcome than *x*, even where *y* and *x* themselves are Pareto-noncomparable.

The term "Kaldor-Hicks efficiency" is often used as a synonym for a potential Pareto improvement test, and that shall be my usage here. Although Kaldor and Hicks actually proposed different versions of such a test, I shall ignore this historical nicety and use the term "Kaldor-Hicks efficiency" to mean the entire genus of potential-Pareto-improvement tests – whatever the precise details.

Kaldor-Hicks tests, whatever their details, focus on the possibility of producing a Pareto improvement via costless, lump-sum redistribution of the total stock of marketed goods[53] associated with a given outcome. Consider two outcomes x= ($c_1 \ldots c_N$, **q**, **p**, **t, b**) and y =(c_1*,…, c_N*, **q***, **p***, **t***, **b***). Vector **q** is the grand vector of marketed goods in *x*: the items that individuals actually consume and directly benefit from. If there are *M* of these, **q** has the form $(q_1^1,...,q_1^M,q_2^1,...,q_2^M,...,q_N^1,...q_N^M)$. In other words, **q** is a particular allocation of each of the *M*

[53] Theoretical treatments of Kaldor-Hicks tests focus on redistribution of goods rather than services, although presumably this could be generalized.

goods among the *N* individuals. Vector **p** is the price vector, listing the prices of the *M* goods. Vector **t** is the technology and stock of inputs in *x*: a description of the factors of production possessed by each producer, as well as the technological processes she employs in transforming those inputs into marketed goods. Vector **b** denotes other individual attributes or background facts. [54] Individual *i*'s consumption c_i, is simply the market value of *i*'s allotment of the *M* goods as per **q**, given the prices in **p**.

One version of the Kaldor-Hicks test looks at possible redistributions of the actual stock of goods in a given outcome. It says: *y* is a potential Pareto improvement over *x* iff there is some eligible reallocation of **q*** (call it **q+**), such that some individuals strictly prefer (**q+**, **b***) to (**q**, **b**), and everyone is at least indifferent. [55] A eligible reallocation is *any* reassignment of the particular allocation of goods to individuals specified by **q* --** as long as the total amount of each good is no more than the total amount of each good in **q***. Thus a reallocation is eligible, for purposes of the Kaldor-Hicks test, even if the actual redistributive mechanism that government would use to try get from **q*** to **q+** would be costly, and would actually produce an allocation with a smaller total stock of goods than **q+**.

This test – like CBA and everything else in welfare economics – is standardly framed in terms of preferences, but we can make the test more generic by framing it in terms of well-being. Thus framed, the test says: given some account of well-being *W*, *y* is a potential Pareto improvement over *x* iff **q+** is some eligible reallocation of **q*** and (**q+**, **b***) is Pareto superior to (**q**, **b**) in terms of *W*.

A variation on this test look at eligible reallocations in both outcomes, saying: *y* is a potential Pareto improvement over *x* iff (1) **q+** is some eligible reallocation of **q*** and (**q+**, **b***) is Pareto superior to (**q**, **b**), and (2) there is *no* **q'** which is an eligible reallocation of **q**, such that (**q'**, **b**) is Pareto superior to (**q***, **b***).

A different version of the Kaldor-Hicks test looks at possible redistributions of the total stock of marketed goods that can be produced in a given outcome, given the productive technology and stock of inputs associated with that outcome. This version of the test (framed generically in terms of well-being) says: given some account of well-being *W*, *y* is a potential Pareto improvement over *x* iff there is some eligible reallocation of the inputs to producers specified by **t***, some production plan which is feasible given the technology specified by **t***, and

[54] Although theoretical work on Kaldor-Hicks tests rarely discuses nonmarket characteristics, CBA certainly does, and so if we're looking to Kaldor-Hicks tests as a potential justification for CBA, we should allow for **b** to include non-market characteristics.

[55] Here and for the next few paragraphs, I drop the price vector, technology and stock of inputs, and vector of individual monetary consumption amounts, $c_1 \ldots c_N$, from the description of the outcomes. This is just to simplify the presentation, and does not at all bear upon my critique of the Kaldor-Hicks tests. As we reallocate goods from **q*** to **q+**, prices will change too; but prices, technology, and inputs are not seen by the proponents of the Kaldor-Hicks test as driving individuals' well-being. Note also that, with outcomes characterized so as to explicitly describe what specific goods individuals consume, consumption, i.e., total expenditure, also does not influence well-being (or at least economists assume it does not).

some eligible reallocation, **q++**, of the vector of goods thus produced, such that (**q++**, **b***) is Pareto superior to (**q**, **b**) in light of *W*.

Here, too, a variation on this test applies it in both outcomes.

Note that the Kaldor-Hicks tests described thus far compare the entire set of eligible reallocations corresponding to one outcome, to the actual allocation achieved in another. A different version compares sets to sets. In other words, it says: *y* is Kaldor-Hicks superior to *x* if, for every eligible reallocation of the marketed goods in *x*, there is some eligible reallocation of the marketed goods in *y* which is Pareto superior. Or it says: for every eligible reallocation of the marketed goods producible in *x* given the technology and stock of productive factors in *x*, there is some eligible reallocation of the marketed goods producible in *y* given the technology and stock of productive factors in *y* which is Pareto superior.

In turn, CBA is very often seen as way to implement one of these Kaldor-Hicks tests. However, Robin Boadway discovered in 1974 that the apparent correspondence between a given version of CBA and a seemingly equivalent Kaldor-Hicks test may be illusory. Specifically, Boadway showed the following. Imagine that we are dealing with what economists call an "exchange economy." There is a single, fixed, total stock of marketed goods in all possible outcomes, and non-market characteristics are ignored. So *x* has the form (c_1, …, c_N, **q**, **p**), where c_i is simply the total market value of the marketed goods and services that *i* uses (as described by **q**), calculated using the extant prices **p**. Outcome *y* has the form (c_1*,…, c_N*, **q*, p***). Finally, let us imagine a case in which *x* and *y* are on the so-called "Pareto frontier" for the exchange economy (the set of allocations of the total fixed stock of goods that are not Pareto inferior to any allocation) -- so that there is no eligible reallocation of the goods in *y* which is Pareto-superior to *x*, and no eligible reallocation of the goods in *x* which is Pareto superior to *x*.

Nonetheless – Boadway showed – it is quite possible, indeed to be expected, that the sum of WTP/WTA amounts for *y* relative to *x* will be positive, and that the sum of WTP/WTA amounts for *x* relative to *y* will be positive. This result is not particularly intuitive, and I will not discuss it in detail here, but the basic point is that the difference between the price vector in *y* and *x* drives a wedge between the CBA test and the Kaldor-Hicks test. Subsequent theoretical scholarship has further elaborated the ways in which various versions of CBA can deviate from various versions of the Kaldor-Hicks test. Boadway and Bruce, surveying this scholarship, observe: "The use of the unweighted sum of household compensating or equivalent variations [i.e., WTP/WTA amounts] as a necessary and sufficient indictor of potential Pareto improvements is rife with difficulties."

However, the most telling objection to the Kaldor-Hicks defense of CBA is different. Even where CBA *does* track the Kaldor-Hicks criterion, it is very hard to see why this feature of CBA provides a *justification* for it.

Imagine that we have used CBA to produce a Pareto-respecting quasi-ordering of some outcome set. Like all the other frameworks considered in this chapter, the justifiability of CBA in this outcome-ranking role will hinge on how it ranks Pareto-noncomparable outcomes. If the reader is content to take the position that all Pareto-noncomparable outcomes are morally incomparable, she doesn't need CBA as a criterion for ordering outcomes. She can simply use the Pareto-quasiordering to do so. Reciprocally, if the reader shares the intuition of this author and many others that the Pareto-quasiordering seems too abstemious -- surely there are *some* cases in which *x* and *y* are Pareto-noncomparable and yet *x* is morally better than, equal to, or worse than *y*, not morally incomparable – she will entertain criteria that extend the Pareto-quasiordering (CBA and the others considered in this chapter) and evaluate those criteria by how they rank pairs of outcomes that are Pareto-noncomparable.

Consider, then, a case in which *x* and *y* are Pareto-noncomparable; some version of CBA (some kind of sum-of-WTP/WTA amounts test) ranks *y* as *better* than *x*; and it turns out that CBA, here, tracks some version of a Kaldor-Hicks test. Outcome *y* is also better than *x* in terms of this Kaldor-Hicks test. There is an eligible reallocation of the stock of marketed goods in *y* which is Pareto superior to *x*; or there is an eligible such reallocation, plus no eligible reallocation of the stock of marketed goods in *x* which is Pareto superior to *y*; or an eligible reallocation of the total stock of marketed goods producible in *y* given local technology and productive factors which is Pareto superior to *x*; or such an eligible reallocation, plus no eligible reallocation of the total stock of marketed goods producible in *x* which is Pareto superior to *y*; or some such similar fact comparing the entire set of eligible reallocations corresponding to *x* and *y*. In general, what these tests do is to map *y* onto a set including *y* plus *stipulated transformations* of *y*, to map *x* onto a set including *x* plus *stipulated transformations* of *x*, and to rank the two outcomes, *y* and *x*, depending on relations of Pareto superiority between one or more members of the *y* set and one or more members of the *x* set.

Why should such a Kaldor-Hicks test convince us that *y* itself is morally better than *x*? How can we leap from a relation of *potential* Pareto superiority to the relation of actual moral betterness? The following line of argument is a complete non sequitur: (1) Outcomes *y* and *x* are Pareto noncomparable; (2) if outcome *w* is Pareto superior to outcome *z*, then *w* is morally better than *z*; (3) *y* and/or some of its stipulated transformations are Pareto superior to *x* and/or some of its stipulated transformations; therefore (4) outcome *y* is morally better than outcome *x*. This is Amartya Sen's objection to the Kaldor-Hicks test:

> In what sense is a rise of 'potential welfare' of interest to *actual* welfare comparisons? Even if gainers *could* overcompensate the losers, why is that an improvement? It might be thought that the answer depends on whether compensations are *actually* paid or not. But there is a problem in *either* case.
>
> …. The particular example, *viz*., the repeal of the Corn Laws in Britain that motivated the formulation of the compensation tests, involved losses for landlords but gains for the rest …. The fact that landlords don't typically receive much sympathy may have played a psychological part in making unpaid compensation more acceptable. But the losers can just as easily be the poorest of the poor. Or, the most deserving

> according to any criterion of desert that we might wish to specify. A change that leaves them losers, though potentially compensatable, may not be an improvement in any obvious sense.
>
> If, on the other hand, compensation is actually paid, then *after* the act of compensation, everyone is at least as well off as before and someone is strictly better off. That being the case, the situation is a welfare improvement on straightforward Paretian grounds. But then no compensation tests are needed, since the Pareto criterion itself is sufficient! Thus it would seem that compensation tests [i.e., Kaldor-Hicks or potential Pareto tests] are either unconvincing (when compensations are not actually made) or redundant (when they are).

In general, the fact that an outcome, action, individual, or other item is *transformable* into an item with some attribute does not mean that the original item is normatively equivalent to one that already has the attribute. For example, it would be absurd to say that an outcome *x* in which individual *i* has a life full of unhappiness and pain yields the same well-being level for him as an outcome *y* in which *i*'s life is happier and more pleasurable – and that *x* and *y* should therefore be treated as morally equivalent -- merely because we could transform the first outcome into the second via a change in *i*'s psychological dispositions. And it would be absurd to say that the appropriate choice by two different actors facing the same, physical choice situation, but with different beliefs, is the same – merely because we could change one actor's belief state into the other by adding some information. There *is* a special kind of case in which we *do* plausibly judge a certain kind of transformability of one item into a second to produce a kind of normative equivalence – namely, where considerations of responsibility are involved. For example, responsibility-sensitive welfarists *might* well plausibly say that an outcome *x* in which *i*'s life is unhappy and painful should be treated as morally equivalent to an outcome *y* in which his life is happier and less painful, if the occurrence of *x* rather than y would be *i*'s responsibility. But the potential Pareto test (in all of its various formulations) builds in no connection to responsibility, as Sen points out: "But the losers can just as easily be … the most deserving according to any criterion of desert that we might wish to specify."

Note that the objection to the Kaldor-Hicks defense of CBA that I have pressed here does not trade on the possibility of interpersonal comparisons. Assume that such comparisons are impossible. Even so, it is very difficult to see how the transformability of *y* into an outcome Pareto-superior to *x* means that *y* itself is better than *x*, given that *y* and *x* are themselves Pareto noncomparable. Nor does the objection trade on exactly how the stipulated variations from *y* and *x* are defined, for purposes of the Kaldor-Hicks test. Imagine that we focus on allocations which are feasible given the administrative costs of reallocation (rather than ignoring such costs, as in the standard versions of the Kaldor-Hicks test). There is a set of vectors of marketed goods including the vector in *y* itself plus vectors which are achievable from *y*, given administrative costs; and a set of vectors of marketed goods including the vector in *x* itself plus vectors which are achievable from *x*, given administrative costs. Some of the elements in the first set bear a relation of Pareto-superiority to some of the elements in the second set. So what? Why does this mean that *y* is better than *x*, since *y* itself is Pareto noncomparable with *x*?

Finally, the objection pressed here does not trade on the well-known fact that CBA and Kaldor-Hicks tests are insensitive to the distribution of consumption or other well-being relevant attributes. Many readers, like the author, will indeed firmly believe that an attractive welfarist choice-evaluation procedure should rank Pareto non-comparable outcomes in a manner that is sensitive to fair distribution. But even if the reader does not share this belief, she should still see the Kaldor-Hicks justification of CBA as unpersuasive for the reasons I have tried to outline.

Parenthetically, one might wonder whether the scholars associated with the Kaldor-Hicks test really saw it as a fundamental principle for ranking outcomes, on a par with other fundamental principles (in particular the Pareto principles). At certain points, at least, some of them suggested a different sort of role for the test – namely as a criterion whose use by government would produce long-run improvement. However, there is a certainly a substantial literature in economics that analyses the Kaldor-Hicks test as a fundamental principle for ranking outcomes. I have argued that it fails in that role. CBA may or may not yield a ranking of an outcome set that corresponds to a Kaldor-Hicks test (given Boadway-type divergences); but even where there *is* such correspondence, that fact does nothing to establish that the ranking achieved by CBA is morally attractive.

CBA as a proxy for overall well-being

In prior work, Eric Posner and I presented a revisionary defense of CBA. We rejected the standard justification for CBA that links it to Kaldor-Hicks efficiency, and we also rejected the view among many economists that interpersonal welfare comparisons are impossible. Instead, we argued for the possibility of such comparisons, and therefore for the intelligibility of the concept of *overall well-being*. And we suggested that CBA was a rough proxy for overall well-being.

Critically, Posner and I did *not* claim that CBA furnished a principle for morally ranking outcomes. We wrote:

> Cost-benefit analysis is not a moral criterion. The fact that the sum-of-[WTP/WTA amounts] is greater in outcome O than in O* does not mean that O is *any way* morally better or more attractive than O*.

Rather, we suggested that CBA was a decision procedure that governmental officials ought to be legally required to employ. In other words, we suggested that a legal structure *L* in which administrative agencies employ CBA as a policy-analysis tool is morally preferable to a legal structure *L** in which they employ some other tool. In general, to endorse a procedure as a component of a morally attractive set of legal institutions is not the same as endorsing the procedure *as a moral decision procedure*. It is not the same as claiming that the procedure works well to furnish guidance concerning which choices, in any given choice situation, are morally better and which are morally worse. (For more on this distinction, see Chapter 8).

Further, we stressed that CBA was at best a *rough* proxy for overall well-being, and that governmental officials should consider using CBA with distributive weights attached, or using a different procedure in policy choice situations characterized by large wealth skews and, thus, large skews in the marginal utility of income.

To see why interpersonal comparability does not strengthen the case for using CBA as a principle for generating a moral ranking of outcomes, let us assume now that the most attractive welfarist decision-evaluation framework includes a theory of well-being, *W*, that generates an interpersonal as well as intrapersonal ranking of life-histories and differences between them. Assume, further, that this ranking is represented by a single utility function $u(.)$, unique up to a positive ratio transformation. (We can therefore refer to the utility assigned to a given life-history as "$ru(.)$", r a positive constant.) Life history $(x; i)$ is at least as good for well-being as life-history $(y; j)$, according to *W*, iff $ru(x; i) \geq ru(y; j)$. The difference between life-histories $(x; i)$ and $(y; j)$ is at least as large as the difference between life-histories $(z; k)$ and $(w; l)$ iff $ru(x; i) - ru(y; j) \geq ru(z; k) - ru(w; l)$.

I assume uniqueness up to a ratio transformation so as to simplify my presentation. My critique of CBA readily extends to the more general case where there is some interpersonal comparability and the ranking of life-histories is represented by a set **U** of utility functions whose elements are not all related by a positive ratio transformation.

In line with the discussion earlier, we are assuming that outcomes have been specified at least in terms of individual consumption and also, perhaps, other characteristics; that WTP/WTA amounts have been given a generic definition in term of *W*, not in terms of actual preference satisfaction; that the connection between well-being and consumption is well-behaved; and that we are ranking outcomes using a "money metric" style sum -of-WTP/WTA test that has been specified to avoid the problems for Pareto superiority created by WTP/WTA reversals. The test says: Pick one baseline outcome *x*; for each individual *i*, determine her WTP/WTA amount for *x*, relative to outcome *y*; add these up across individuals; and assign the negative of the sum to *y*. This number is the "monetized aggregate benefit" of *y*. Then one outcome is at least as good as a second iff its monetized aggregate benefit is at least as large.

Formally, if $x = (c_1,\ldots, c_N, \mathbf{b})$, and $y = (c_1^*,\ldots, c_N^*, \mathbf{b^*})$, then individual *i*'s WTP/WTA amount for *x*, relative to *y*, is the amount Δc_i^y such that life-history $(c_1,\ldots, c_i - \Delta c_i^y, \ldots, c_N, \mathbf{b}; i)$ is ranked by *W* as equally good as life-history $(c_1^*,\ldots,c_N^*, \mathbf{b^*}; i)$. Because the well-being assigned to life-histories is now representable by a utility function, we can express WTP/WTA amounts in terms of the utility function. Δc_i^y is such that $ru_i(c_1,\ldots, c_i - \Delta c_i^y, \ldots, c_N, \mathbf{b}) = ru_i(c_1^*,\ldots,c_N^*, \mathbf{b^*})$. And the monetized aggregate benefits rule for ranking outcomes is: outcome *y* is at least as good as outcome *z* iff $\sum_{i=1}^{N}(-\Delta c_i^y) \geq \sum_{i=1}^{N}(-\Delta c_i^z)$.

It is clear that the ranking of outcomes in terms of their monetized aggregate benefits need not match their ranking in terms of overall well-being. (This is why Posner and I argued that CBA is at most a *rough* proxy for overall welfare.) The ranking of outcomes in terms of overall well-being is given by a *utilitarian* SWF. On the assumption that utility is unique up to a positive ratio transformation, the utilitarian SWF says: outcome *y* is at least as good as outcome *z* iff $\sum_{i=1}^{N} ru_i(y) \geq \sum_{i=1}^{N} ru_i(z)$.

To see why the CBA test and the utilitarian test can deviate, consider first the simplest case, where outcomes are specified solely in terms of consumption and each individual's utility is just a function of his own consumption. Consider, now, a graph relating an individual's consumption level in some outcome to his *utility*. There is no particular reason to assume that this graph will be linear. Indeed, economists and others who accept the possibility of interpersonal welfare comparisons typically assume that money has "declining marginal utility." In other words – this is just what it means for money to have declining marginal utility – they assume, very plausibly, that this sort of graph will be strictly *concave*.

[chart]

But, given a strictly concave graph, or indeed any non-linear form, the monetized aggregate benefits test can deviate from the utilitarian SWF.

The point readily generalizes to more complicated cases. Imagine now that outcomes are characterized both in terms of individual consumption as well as other individual attributes and/or background facts, summarized in the term **b**. An individual's utility is now a function both of her consumption plus these other attributes and/or facts. It takes the form $ru(c_i; \mathbf{b})$. *If* this utility function can, in turn, be additively decomposed into the "subutility" of individual consumption" plus a "subutility" value assigned to the **b** term, and *if* the subutility of consumption is linear rather than concave (or nonlinear in some other way), then the aggregate net benefits test and the utilitarian SWF will necessarily coincide. But these are very restrictive conditions on the individual utility function. It is implausible to think that a utility function tracking individual well-being will necessarily satisfy these conditions. Once they are relaxed, the monetized aggregate benefits test and the utilitarian SWF can deviate, as the following table illustrates.

> If the utility of a life-history with consumption *c* and other characteristics **b** has the special additive form $ru(c, \mathbf{b}) = r(h(c) + v(\mathbf{b}))$, where $h(c) = kc+l$, *k* and *l* constants, then the monetized aggregate benefits test and the utilitarian SWF will coincide. One simple functional form that deviates from this form is $ru(c, \mathbf{b}) = r(h(c) + v(\mathbf{b}))$, where $h(c)$ is nonlinear. For example, $h(.)$ might be strictly concave, so that consumption has declining marginal utility. Another simple functional form which is different from the special additive form is multiplicative, i.e., $ru(c, \mathbf{b}) = r(cv(\mathbf{b}))$. The following examples show how both of these simple non-additive utility functions can produce deviation between the utilitarian SWF and the monetized aggregate benefits test [Table]

Moreover, even if CBA *were* a perfect proxy for overall well-being, non-utilitarian welfarists who believe that the most attractive moral ranking of outcomes can *deviate* from overall well-being would still properly reject CBA as an outcome-ranking principle. Consider, in particular, the continuous prioritarian SWF: the SWF that I will defend in Chapter 4, as against other distribution-sensitive SWFs, and against the utilitarian SWF. In the case where utility is unique up to a ratio transformation, and the continuous prioritarian SWF is the "Atkinsonian" variety,[56] the criterion for ranking outcomes is: outcome *y* is at least as good as outcome *z* iff $\frac{1}{1-\gamma}\sum_{i=1}^{N} r^{1-\gamma}u_i(y)^{1-\gamma} \geq \frac{1}{1-\gamma}\sum_{i=1}^{N} r^{1-\gamma}u_i(z)^{1-\gamma}$, where γ is an inequality aversion parameter, a positive number that encapsulates the degree of priority given to well-being changes affecting worse-off individuals. Assume, now, that individual utility has the special additive form described in the preceding paragraph, so that the monetized aggregate benefits test and the utilitarian SWF coincide. Because the *prioritarian* SWF is distribution-sensitive and can deviate from the utilitarian SWF, it can also – therefore – deviate from the monetized aggregate benefits test.

[Table]

So let us now turn to the possibility of CBA with distributive weights. There is actually a substantial academic literature on this topic. The literature typically draws a close connection between distributively weighted CBA and social welfare functions, suggesting that distributive weights be specified using some social welfare function. Although CBA, in governmental practice, is typically undertaken without distributive weights – certainly that is true in the United States, which has spearheaded the use of CBA as a governmental decision procedure – there are some exceptions. Distributive weights have been used at the World Bank (although apparently with less frequency in recent years), and the current guidance document for CBA in the U.K. encourages their use.

Distributively weighted CBA takes the form of summing individual WTP/WTA amounts multiplied by individual weights. Rather than assigning each outcome *y* a monetized aggregate benefits value equaling $\sum_{i=1}^{N}(-\Delta c_i^y)$, and ranking outcomes in the order of these values, distributively weighted CBA assigns each outcome a monetized *weighted* aggregate benefits value equaling $\sum_{i=1}^{N} w_i^y(-\Delta c_i^y)$, where w_i^y is the weight for individual *i* in outcome *y*, and ranks outcomes corresponding to *these* values.

[56] I argue for the Atkinsonian SWF because it is the only continuous prioritarian SWF which is invariant to a ratio transformation – and this feature is indeed useful for purposes of the discussion, addressing distributively weighted CBA in the simple case where **U** has the form *ru*(.).

It is easy to see that, in principle, distributive weights *can* cure the deviation between CBA and various SWFs. I am assuming, here, that the weighting scheme has a fair degree of flexibility: the weighting factor for an individual can vary from outcome to outcome, and can be a function of a variety of well-being relevant characteristics, not just consumption. [57]

Using a flexible weighting scheme of this sort, it is trivial to specify weights so as to cure any deviation between CBA and the utilitarian SWF. Simply calculate the weighting factor w_i^y for individual *i* in outcome *y* by taking the difference between *i*'s utility in the baseline outcome *x* and *y*, and dividing by the WTP/WTA amount Δc_i^y. [58] Then it is absolutely straightforward to see that the ordering of outcomes using the values assigned them by the formula $\sum_{i=1}^{N} w_i^y(-\Delta c_i^y)$ is exactly the same as the ordering using the values assigned by the utilitarian SWF, i.e., $\sum_{i=1}^{N} ru_i(y)$. Similarly, using a flexible weighting scheme, it is trivial to specify a different set of weights so as to cure any deviation between CBA and the prioritarian SWF of the Atkinsonian variety. Now, calculate the weighting factor w_i^y for individual *i* in outcome *y* by taking the difference between *i*'s *transformed* utility in the baseline outcome *x* and her *transformed* utility in outcome *y*, and dividing this difference by the WTP/WTA amount Δc_i^y. [59] With the individual weights calculated in this manner, it is straightforward to see that the ordering of outcomes using the values assigned them by the formula $\sum_{i=1}^{N} w_i^y(-\Delta c_i^y)$ is exactly the same as the ordering using the values assigned to outcomes by the Atkinsonian SWF, i.e., $\frac{1}{1-\gamma}\sum_{i=1}^{N} r^{1-\gamma}u_i(y)^{1-\gamma}$

[57] Proponents of distributive weights often discuss a simpler approach, where an individual is assigned a single weight as a function of his attributes just in the baseline outcome (perhaps just his consumption). Even where outcomes are just specified in terms of consumption, distributively weighted CBA using a single individual weight (rather than an outcome-specific individual weight) will not reliably mirror an SWF when an individual's consumption levels in various outcomes are substantially different. And where utility is a function of consumption as well as other attributes, distributive weights will not reliably cure the deviation between CBA and an SWF if weights are solely a function of individual consumption. Another possible way to specify distributive weights in this last case is to *normalize* consumption to reflect non-consumption attributes, and make distributive weights a function of normalized consumption plus the reference level of the non-consumption attributes. On normalization, see below, this chapter, and Chapter 6.

[58] More precisely pick any *u*(.) arbitrarily, and set $w_i^y = (u_i(x) - u_i(y))/\Delta c_i^y$. Let us denote $\sum_{i=1}^{N} u_i(x)$ as K. Then $\sum_{i=1}^{N} w_i^y(-\Delta c_i^y) = \sum_{i=1}^{N}(u_i(y) - u_i(x)) = \sum_{i=1}^{N} u_i(y) - K$. Thus the ordering of outcomes achieved by ranking outcomes according to their monetized weighted aggregate benefits is exactly the same as ordering them by summing utilities using *u*(.). And this is, in turn, the same for all *u*(.) in **U** if that set is unique up to a positive ratio transformation.

[59] More precisely, pick one *u*(.) arbitrarily and set $w_i^y = [\frac{1}{1-\gamma}u_i(x)^{1-\gamma} - \frac{1}{1-\gamma}u_i(y)^{1-\gamma}]/\Delta c_i^y$.

Moreover, this procedure for curing the deviation between CBA and utilitarian or prioritarian SWFs generalizes to the case in which utility is not necessarily unique up to a positive ratio transformation. [60] It generalizes beyond the Atkinson SWF to any continuous prioritarian SWF. Indeed, it generalize to any SWF that can be encapsulated in a mathematical function, i.e., to any SWF that takes the form: outcome x is at least as good as outcome y iff, for all $u(.)$ belonging to **U**, $w(u(x))\geq w(u(y))$.

However, it is unclear why the welfarist who accepts the possibility of interpersonal comparisons and the representability of well-being via utility numbers should advocate using distributively weighted CBA as the outcome-ranking component of her choice-evaluation procedure. Why not rank outcomes directly, using whatever SWF the welfarist takes to be attractive, rather than via this more roundabout technique? Distributively weighted CBA, like the direct application of an SWF, requires the estimation of a utility function – which is critical in specifying weights. And in the case of a non-utilitarian SWF, it also requires specifying the parameters of the $w(.)$ function – for example, the inequality aversion parameter γ for the Atkinson SWF. But, in addition, it requires the extra step of converting well-being impacts into WTP/WTA amounts. Why takes this extra step? Why not more directly rank outcomes using the formula: outcome x is at least as good as outcome y iff, for all $u(.)$ belonging to **U**, $w(u(x))\geq w(u(y))$?

One possible answer has to do with optimal legal structures. Assume it is indeed morally attractive to establish legal institutions whereby some range of decisionmakers, within some given society's government, are legally required to employ an SWF of some kind in evaluating their decisions. If CBA is currently widely used within the government or society, then – as a kind of transitional measure – it may be perhaps appropriate to instruct those decisionmakers to build upon their expertise with CBA and attach distributive weights to WTP/WTA amounts, rather than shifting to direct implementation of an SWF. But this line of argument, which concerns the content of certain legal requirements, does not address the distinct (and logically prior) question now on the table: What is the most attractive way to outfit a moral decision procedure? Why does the welfarist have good reason to favor the blueprint of distributively weighted CBA, rather than the SWF blueprint, as the general format for ranking outcomes as morally better or worse?

In any event, even if it *is* the case that distributively weighted CBA is an attractive format for a moral choice-evaluation procedure, this proposition does not undermine the basic claim of this chapter. I claim that existing policy-evaluation approaches are either less attractive than the SWF approach, as a basis for morally ranking outcomes, or alternatively are variations on the SWF approach. By a variation on the SWF approach, I mean an approach to ranking outcomes that accepts that well-being is interpersonally comparable to some extent; accepts that well-being

[60] In that case, we calculate WTP/WTA amounts and weights for each $u(.)$ belonging to **U**; assign each outcome a monetized weighted aggregate benefit for each $u(.)$; and order outcomes depending on the entire set of monetized weighted aggregate benefit amounts assigned each outcome.

is measurable by utilities; and uses these utilities to generate a ranking of outcomes which is identical to the ranking generated by some Paretian, anonymous SWF. And this is precisely what distributively weighted CBA is.[61]

Other roles for CBA

This section has thus far focused on whether CBA provides an attractive welfarist criterion for morally ordering outcomes. Without distributive weights attached, CBA fails in this role, or so I have argued. Distributively weighted CBA, too, seems less attractive than ranking outcomes via direct application of an SWF, and in event is simply a variation on the SWF approach.

However, nothing in the analysis up to this point precludes CBA from playing a function other than as a criterion of the moral goodness of outcomes. I will revert, now, to using "CBA" to mean some version of a sum-of-WTP/WTA test without distributive weights.

Assume that the most attractive welfarist decision procedure uses a SWF of some kind (*not* CBA) as the outcome ranking principle. This SWF will say: x is at least as good as y iff R ranks $u(x)$ as being at least as good as $u(y)$, for all $u(.)$ belonging to **U**, where R is a Paretian and anonymous rule for ranking utility vectors.

This decision procedure, besides telling us how to construct an outcome set and rank outcomes, will also provide some principles (whatever they might be) for constructing a choice set **A**, and for ranking choices in light of the probability distribution over outcomes each choice yields. Because the decisionmaker is bounded in her cognitive abilities, **A** too will be bounded. It will not include *every* action that the decisionmaker is able to perform; thinking about such a choice set would be cognitively impossible.

CBA might function, here, as a *choice-set- expansion heuristic*. To see how this might work, let us imagine (to start) that the decisionmaker is operating under conditions of certainty. Each choice in **A** yields one outcome for sure. Let us designate by a^x the action that yields outcome x with certainty, and so forth for other outcomes. Imagine, now, that the decisionmaker has used the SWF formula to rank the outcomes corresponding to the choices in **A,** and that she has determined that y is the best outcome produced by some choice in **A** and that a^y is therefore the best action. She is about to select a^y -- until someone points out that a^z, which is also in **A**, is preferred to a^y by CBA, and that CBA here tracks a simple kind of Kaldor-Hicks test. In other words, there is some reallocation of the consumption goods in z, call that outcome z^*, which is Pareto superior to y.

This observation should prompt the decisionmaker to consider expanding her choice set. Might there be some new choice b, which is not in the initial choice set, but which the

[61] Nor is seeing distributively weighted CBA this way artificial, since as already mentioned those who favor it *do* generally take the position that weights should be set with reference to a social welfare function.

decisionmaker is physically capable of performing, and that yields either z^* or at least some other outcome which is Pareto superior to y? There need not be such a choice b, given the administrative costs of the tax system, the distortionary effects of the income tax, the limited ability of the tax authorities to detect individuals' abilities, and so forth. The extant technology of taxation might not make it possible for the decisionmaker to "vary" a^z, so that consumption is shifted, without too much loss in well-being, and an outcome Pareto superior to y is produced. Further, whatever the technology of taxation, the decisionmaker may not be able to use the tax system to redistribute consumption – either because she is not a governmental decisionmaker, or because she is an official within some non-tax governmental body.

Still, if a feasible choice b did exist, the SWF formula – x at least as good as y iff R ranks $u(x)$ as being at least as good as $u(y)$, for all $u(.)$ belonging to **U** -- would *favor b* over the initially chosen action a^y. This is because the SWF is Paretian: it always prefers a Pareto-superior to a Pareto-inferior outcome.

It is critical to be clear about the role that CBA is playing here. If decisionmakers were unboundedly rational, a moral decision procedure would not need choice-set-expansion heuristics: the decisionmaker would simply be instructed, at the outset, to consider all possible alternatives. However, given bounded rationality, an attractive moral decision procedure might plausibly take the form of coupling guidance for the construction of an initial choice set with guidance to "search" for additional choices under certain conditions. Further, because the SWF's ranking of outcomes is *Paretian*, the fact that the outcome z yielded by a choice within the initial choice set, a^z, is *potentially Pareto superior* to the outcome y yielded by the favored choice a^y, is plausibly an attractive *trigger* for a search for an additional option b that will yield an outcome Pareto superior to y. Crucially, however, the *ranking* of outcomes, for purposes of choice from whatever choice set is at hand (the initial choice set, or the choice set expanded after search), is performed by the SWF, not by CBA.

I am not actually arguing for the use of CBA as a choice-set-expansion heuristic. Note that the construction of a CBA test that will usefully function as such a heuristic becomes more complicated in the (realistic) case of choice under uncertainty. Rather, I am trying to clarify that this sort of role for CBA is distinct from its use as an outcome-ranking principle; that nothing in my argument against CBA as an outcome-ranking principle precludes this role; and that, indeed, if we use a Paretian SWF as the outcome-ranking principle, there is some substantial plausibility in thinking that CBA should play this role as a choice-set-expansion heuristic.

Actually, there is already a scholarly literature that uses CBA as something like a choice-set-expansion heuristic. This literature, which will be discussed in Chapter 8, works *within* the optimal tax tradition. It uses a Paretian, anonymous SWF, *not* CBA or a Kaldor-Hicks standard, as the fundamental criterion for evaluating different tax and non-tax policies. But it suggests that, at least under certain behavioral assumptions about the incentive effects of the income tax, a governmental regulation or other non-tax policy that fails a cost-benefit test can always be

improved upon (in light of the SWF) by choosing a different policy and coupling that with some change to the schedule of income taxes.

A second possible role for CBA, also suggested by this literature and discussed in Chapter 8, has to do with morally optimal legal structures. As between a legal structure *L* that instructs non-tax decisionmakers to use CBA as their policy analysis tool, and a legal structure *L** that instructs non-tax decisionmakers to use a different tool, it *might* be the case that a Paretian SWF would actually favor *L* over *L**. Perhaps legal structures that channel redistribution through the tax system and tell non-tax decisionmakers to use a simple-cost-benefit text would make everyone better off. A different sort of argument for *L* rather than *L**, also considered in Chapter 8, is that the benefits of cost-benefit justified policies would be "spread around," so that everyone would benefit from *L* rather than *L** in the long run (and thus that a Paretian SWF would favor *L* over *L** even without a tax system that converts potential into actual Pareto improvements). Posner and I, as mentioned, have also argued for the moral attractiveness of a legal system that instructs decisionmakers to use CBA as a policy analysis tool.

These claims about the role of CBA in optimal legal structures, whatever their ultimate persuasiveness, are again conceptually distinct from the claim criticized in this chapter -- that CBA is a principle for morally ranking outcomes. Unfortunately, much of the discussion of CBA and the Kaldor-Hicks principle has been marred by a failure to clearly delineate what particular function is being mooted.

[The remainder of chapter 2, which is not included here, discusses inequality metrics, other equity metrics, and cost-effectiveness analysis.]

Chapter 3: Well-Being and Interpersonal Comparisons

A welfarist decision procedure requires an account of well-being. The SWF approach requires not just any such account, but one that allows for interpersonal comparisons of well-being, and that is representable by a set **U** of utility numbers.

But what *is* the most attractive account of well-being? Why believe that it allows for interpersonal comparisons? What would such comparisons consist in? And even if interpersonal comparisons *are* possible, how do we construct numerical utilities that represent the well-being ranking of life histories or the well-being differences between life histories?

This chapter grapples with these issues. Drawing from the philosophical literature, I offer an analysis of well-being in terms of fully-informed, fully-rational, extended preferences. This account *is* preferentialist, for reasons I will elaborate, but it does not tie well-being to preferences in a crude or simplistic way. Rather, it is designed to be sensitive to important platitudes about well-being – that well-being has critical force, and that it cannot be too "remote" from the subject – which have been part of the motivation for philosophical views that draw a nexus between well-being and mental states or objective goods, rather than reducing an individual's well-being to the satisfaction of her actual preferences.

The idea of an *extended preference* derives from John Harsanyi. This idea fits hand-in-glove with the basic architecture of a welfarist decision procedure, built around a set of life-histories deriving from an outcome set and a group of N individuals. An individual's ordinary preferences take the form of ranking outcomes and choices. Individual k prefers outcome x to outcome y, or choice a to choice b. By contrast, an individual's *extended* preferences take the form of ranking *life-histories*. Individual k has an extended preference for life-history $(x; i)$ over $(y; i)$.

Of course, what it means to have an extended-preference is a subtle question. But let us assume that individuals *can* have extended preferences – indeed, not just rankings of life-histories, but rankings of lotteries over life-histories. This is what Harsanyi assumes, and I believe the assumption can be defended. Harsanyi then proposes to use expected utility (EU) theory to construct utility functions that represent individuals' extended preferences, and this shall be my approach as well. Each individual k in the population, if fully informed and rational, would rank the set of lotteries over life-histories consistent with EU theory.[62] This individual ranking can be represented by an individual set of utility functions, $\mathbf{U}^k$. (More precisely, the set $\mathbf{U}^k$ represents individual k's ranking of lotteries over life-histories, *and* is "zeroed out" so as to assign zero to nonexistence.). Pooling these individual sets across all N individuals, we arrive at the set **U**. The set **U**, thus created, will not simply represent intra- and

[62] More precisely, as I shall explain, each individual k ranks different subsets of life-history lotteries, subsets in which all the life-histories involved concern a single subject. Individual k doesn't directly rank all life-history lotteries, but rather constructs $\mathbf{U}^k$ via her rankings of these subsets plus certain other information.

interpersonal comparisons of well-being levels. It will also represent intra- and interpersonal comparisons of well-being *differences*, and, finally, will allow us to make statements about the ratios between the well-being levels of different life histories.

The chapter begins by surveying the philosophical literature on well-being. Some philosophers link well-being to preference satisfaction; some, to mental states; some, to "objective goods"; and some offer hybrid accounts which combine these elements. This diversity of philosophical approaches is mirrored in economics and in other literatures where the concept of well-being plays a role. A preference-satisfaction account of well-being has, of course, long been the dominant view within economics. However, much scholarship in economics now focuses on the determinants of "happiness" or "subjective well-being." Many economists in this area, as well as psychologists engaged in closely related work, stress the importance of positive mental states and the avoidance of negative ones for human welfare. Finally, Amartya Sen's scholarship on "capabilities" has triggered much recent work in development economics that is skeptical of both preferentialist and mental-state views of well-being, and that seems closer to an objective-good view. And various other bodies of work, for example the literature on "social indicators," also trend towards an objective-good approach.

What accounts for this diversity and controversy? The answer, I suggest, is both substantive and metaethical. As a substantive matter, an attractive account of well-being should be sensitive to various truisms about well-being, several of which I have already mentioned: that well-being has critical force; that it is not too "remote" from the individual involved; and finally that it has motivational force. Developing an account which is true to these platitudes and which is consistent with our intuitions about well-being in particular cases has proved difficult indeed. Adding further fuel to the dialectical fire is a dispute about whether the well-being associated with different lives is ultimately grounded in value-free preferences for such lives, or whether instead well-formed preferences are always "value laden," the result of judgments or perceptions of well-being. The contending positions in this dispute flow naturally from competing metaethical views – in particular, from competing variants of an "ideal approval" account of moral facts.

The chapter next discusses the problem of interpersonal comparisons. While the nature of well-being is philosophically controversial, the possibility of interpersonal comparisons has not been much disputed by philosophers. Economics presents a mirror image. Economists (at least until the recent rise of "happiness" scholarship) have tended to concur in analyzing well-being in terms of preference satisfaction. But while some economists (particularly those who use social welfare functions) accept the possibility of interpersonal comparisons, many others do not. In this section of the chapter, I present a generic case for interpersonal level and difference comparisons, and then describe Harsanyi's use of extended preferences and EU theory to make sense of such comparisons.

Having reviewed both the philosophical literature and the question of interpersonal comparisons, the chapter presents my own account of well-being, which I have already adumbrated: one that derives interpersonal level and difference comparisons from a set **U** of utility functions that represent the fully-informed, fully rational, extended preferences of the various individuals in the population regarding life-histories, lotteries over life histories, and comparisons to nonexistence.

Although my account of well-being is indebted to Harsanyi's work, it differs from his account in a number of critical respects. Harsanyi never offers a clear account of how an extended preference is possible. I grapple with that problem at length. Further, unlike Harsanyi, I do not presuppose that individuals have identical extended preferences. While Harsanyi argues for utilitarianism, I will *reject* utilitarianism and instead will use the set **U** as the input to a prioritarian SWF. Unlike Harsanyi, I recognize the need to provide a substantive defense of the use of EU theory to measure well-being differences. Finally, although Harsanyi accepts that individuals' preferences may need to be laundered to some extent, he ignores the critical point that extended preferences need to be defined so as to "screen out" features of outcomes that are too remote from the subject's well-being.

One terminological point bears mention at the outset. Preferences, whether ordinary or extended, can be "strict" or "weak." To strictly prefer one item to another entails being in favor of the first – ranking it higher. To weakly prefer one item to another means strictly preferring the first or being indifferent between the two. Throughout the chapter, whenever I characterize someone as "preferring" one item to another (whether this is an ordinary preference regarding choices and outcomes, or an extended preference regarding life histories), I mean a strict preference, except where I specifically characterize the preference as weak.

What is Well-Being? Philosophical Debates

Well-being, at least for purposes of this book, is a *normative* concept. It figures, here, within a normative (in particular, a moral) theory, namely welfarism.

How is the normative concept of well-being best understood? This topic has been most rigorously addressed by philosophers. This section reviews the range of accounts of well-being proposed within the philosophical literature; shows how this diversity is mirrored in other scholarly fields where the normative concept of well-being plays a role; and suggest that diversity flows both from the first-order difficulty of reaching a point of reflective equilibrium regarding the nature of well-being, and from metaethical disputes.

In this section, I present and analyze philosophical disputes about the nature of well-being by focusing on what different philosophical schools say about the content of *intra*personal comparisons – namely, the conditions under which life-history (x; i) is better for individual i than life-history (y; i). These differences, obviously, carry over to disagreements about the content of interpersonal comparisons. However, the very possibility of interpersonal comparisons raises

further and distinct difficulties, engaged in the next section. So as to avoid muddying the waters, this section generally focuses on the intrapersonal case.

Preferences, Mental States, and Objective Goods: Three Themes in the Philosophical Literature on Well-Being

Derek Parfit, in *Reasons and Persons*, writes that there are three plausible theories of well-being: a desire-fulfillment theory, a hedonistic theory, and an objective list theory. Others have drawn similar taxonomies. I suggest that it is most useful to think of desire (preferences), hedonic states (or, more generally, good and bad mental states), and objective goods as different possible *elements* of an account of well-being -- rather than as the defining features of mutually exclusive accounts. A nuanced classification of theories of well-being should allow for hybrid accounts. For example, well-being might be analyzed in terms of preferences for mental states. Nor should we be too rigid in differentiating objective goods and preferences. For example, one plausible understanding of "objective goods" is that these are the items which individuals, under ideal conditions, converge in self-interestedly preferring.

Let us start with preferences. A preference is a *choice-relevant ranking*: a ranking of items of some sort, on the part of some person, that disposes her to make choices. A variety of kinds of items can be ranked by preferences. At a minimum, an individual can have preferences regarding *choices* (possible actions) and *outcomes*.

What it means to have a preference (a choice-relevant ranking) regarding choices is pretty straightforward: if I prefer action *a* to *b*, then I am disposed to choose *a* rather than *b* when both are on offer. In the case of preferences regarding outcomes, what this means is: I have an ordering of outcomes which, together with my beliefs about which outcomes different choices will yield, disposes me to make different choices.

That individuals can have preferences for both choices and outcomes is uncontroversial among philosophers and also among welfare economists. [63] These are what I shall term "ordinary" preferences. Later in the chapter, I will argue that an individual can have a preference regarding a third type of item – namely, a preference regarding a life-history. Preferences regarding life-histories – extended preferences -- will be vital to explicating interpersonal comparisons. However, the concept of an extended preference does not figure much in the philosophical literature on well-being, and will not be discussed here. Insofar as this section discusses the various linkages that philosophers draw between preferences and well-being, I mean preferences for outcomes or preferences for choices.

[63] Some welfare economists adopt the problematic view that an individual's choices perfectly reveal her preferences over outcomes – but even these economists do not deny that the concept of a preference regarding an outcome ("state") is a meaningful one.

It is standard in the literature to distinguish between an individual's actual preferences and her idealized preferences – between what the individual is in fact disposed to choose, and what she *would* be disposed to choose were she to have more information, be in a calm and deliberate state, and so forth. One subtlety here, is that the very ascription of actual preferences to an individual entails her satisfying minimal rationality conditions. For example, imagine that Jim has an attitude of some sort regarding outcomes, connected to his choices, but this attitude fails a condition of asymmetry and fails a condition of transitivity. Jim has this attitude regarding the ordered pair of outcomes (x, y), but also regarding the ordered pair of outcomes (y, x); and although he has the attitude regarding the ordered pairs of outcomes (x, y) and (y, z), he lacks the attitude regarding the ordered pair of outcomes (x, z). Jim's attitude, presumably, is not a preference: because not asymmetric, it is not the attitude of strict preference; because intransitive, it is not the attitude of strict preference, weak preference, or indifference.

Notwithstanding this subtlety, the distinction between actual preferences and idealized preferences remains sensible. The minimal rationality conditions required even to have an actual preference are weak – so that it is perfectly intelligible to distinguish between an individual's actual preferences, and idealized, counterfactual preferences that meet stronger conditions of fuller information, fuller rationality, and so forth.

That the well-being realized by an individual in different outcomes depends on her preferences over those outcomes is a view adopted, not just by welfare economists, but by a distinguished tradition within moral philosophy. In his hugely influential early twentieth century work, *The Method of Ethics*, Henry Sidgwick had suggested that "a man's future good on the whole is what he would now desire and seek on the whole if all the consequences of all the different lines of conduct open to him were accurately foreseen and adequately realized in imagination at the present point in time." Rawls' *Theory of Justice*, drawing from Sidgwick, provides a preference-based analysis of an individual's lifetime well-being. The best life-history for a person is the one that accords with a rational plan of life, namely "the plan that would be decided upon as the outcome of careful reflection in which the [individual] reviewed, in light of all the relevant facts, what it would be like to carry out [different possible plans] and thereby ascertained the course of action that would best realize his more fundamental desires." Richard Brandt argues that what is rational for an individual consists in the satisfaction of those of his desires that would survive a process which Brandt terms "cognitive psychotherapy."

> This whole process of confronting desires with relevant information, by repeatedly representing it, in an ideally vivid way, and at an appropriate time, I call *cognitive psychotherapy*. I call it so because the process relies simply upon reflection on available information, without influence by prestige of someone, use of evaluative language, extrinsic reward or punishment, or use of artificially induced feeling-states like relaxation. It is *value-free reflection*.

Brandt goes on to suggest that an individual's well-being might consist in what the individual self-interestedly prefers after "cognitive psychotherapy."

One difficulty with full-information preferentialist accounts of well-being is that what is good for us may change as we become better informed. Drinking vintage wine rather than the ordinary stuff may only be better for me if I drink it with sufficient knowledge to be able to appreciate its quality. Peter Railton, sensitive to this difficulty, suggests that the well-being of an individual (whatever his actual informational state) is determined by what his idealized, fully informed self (taking account of the individual's actual informational state) would want him to want.

> [A]individual's good consists in what he would want himself to want, were he to contemplate his present situation from a standpoint fully and vividly informed about himself and his circumstances, and entirely free of cognitive error or lapses of instrumental rationality.

As these quotations suggest, philosophers who analyze well-being in terms of preference-satisfaction tend to build in strong idealization conditions, requiring that the relevant preferences be very fully informed and meet high standards of rational deliberation. This is an important difference between preferentialist strains within welfare economics and philosophical scholarship about well-being. Welfare economists require that preferences satisfy basic rationality conditions – the traditional such conditions within economics are that an individual's preferences regarding outcome should be a complete quasiordering, and that her preferences regarding choices should comply with EU theory – but economists typically do not assert that preferences relevant to well-being are idealized in any more robust sense.

A second theme that occurs repeatedly within the philosophical literature is the connection between well-being and good or bad mental states. Jeremy Bentham, famously, saw well-being as the occurrence of pleasurable mental states and the avoidance of painful mental states. A "pleasure," for Bentham, was a mental state with the property of feeling good; a "pain," a mental state with the property of feeling bad. Other philosophers have concurred with Bentham in linking well-being to mental states without sharing his view that the welfare-relevant attribute of a mental state consists in how it feels. John Stuart Mill, for example, analyzed well-being in terms of the occurrence of *desirable* mental states. The contemporary philosopher Fred Feldman defends an account of well-being that he terms "attitudinal hedonism." He distinguishes between "sensory pleasures" and "attitudinal pleasures" and analyzes well-being in terms of the latter.

> A person takes attitudinal pleasure in terms of some state of affairs if he enjoys it, is pleased about it, is glad that it is happening, is delighted by it. So, for example, suppose that you are a peace-loving person. Suppose you take note of the fact that there are no wars going on. The world is at peace. Suppose you are pleased about this. You are glad that the world is at peace. Then you have taken attitudinal pleasure in a certain fact – the fact that the world is at peace. Attitudinal pleasures are always directed onto objects, just as beliefs and hopes and fears are directed onto objects. This is one respect in which they are different from sensory pleasures. Another difference is that attitudinal pleasures need not have any 'feel.' We know we have them not by sensation, but in the same way (whatever it may be) that we know when we believe something, or hope for it, or fear that it might happen.

Although Feldman does in fact describe his view of well-being as a "hedonistic" view, that term is potentially misleading, because it suggests that mental states make a difference to well-being in virtue of how they feel– which clearly is *not* Feldman's view. A more helpful way of describing what Bentham and Mill have in common with each other, and with a number of contemporary philosophers of well-being (including not only Feldman, but also Wayne Sumner, Mark Bernstein, and Roger Crisp), is that the account of well-being proposed by each of these philosophers focuses on the occurrence or nonoccurrence of a certain kind or kinds of mental states as the basic source of well-being. This more generic characterization of such accounts encompasses the full range of plausible views about what the well-being relevant properties of mental states are.

The reader might, quite understandably, find it confusing that I distinguish between "preferences" and "mental states" as two distinct types of elements in accounts of well-being. Isn't a preference a mental state? For our purposes – namely, seeing what different philosophical views suggest about how to rank life-histories – the difference between a "preferentialist" and "mental state" approach can be framed as follows. What is it about life-history (x; i) that makes it better, worse, or equally good for i's well-being as life-history (y; i)? One type of answer, the preferentialist answer, points to i's preferences as between the outcomes that figure in these life-histories.[64] It says: (x; i) is better than (y; i) because i prefers outcome x to outcome y; or would prefer x after cognitive psychotherapy; or i's idealized counterpart would prefer outcome x; or something like that. A different type of answer, the mental state answer, points to the mental states that i experiences in x and y. It says: (x; i) is better than (y; i) because, in some respect, individual i's mental states in x are different from his mental states in y.

A further subtlety, which the discussion has already illustrated, is that a given account of well-being can rely upon *both* these elements: it can make *both* preference satisfaction, *and* the occurrence of a certain type of mental state, a basis for the well-being ranking of (x; i) as against (y; i). Consider a view which says that (x; i) is at least as good for i's well-being as (y; i) iff i weakly prefers his mental states in x to his mental states in y in virtue of the intrinsic properties of his mental states in both outcomes.[65] Such an approach *does* make the difference between (x;

[64] Another kind of answer, also preferentialist, becomes possible once we have the idea of extended preferences in hand. It makes the ranking of (x; i) and (y; i) depend on everyone's extended preferences between the two life-histories. The account of well-being I defend later in the chapter adopts this answer.

[65] We can classify mental states in terms of their intrinsic properties or in terms of their relational properties. Intrinsically, a true belief and a false belief with the very same content are the same; relationally, they're not. The intrinsic properties of mental states are a matter of what's "in the head." (More precisely, the intrinsic property of a mental state is (1) an intrinsic property of the person which (2) is a mentalistic property. Arguably, it's not strictly correct to ascribe properties to a mental state, which is itself a property of a person rather than a separate entity.) A theory that makes the well-being ranking of (x; i) and (y; i) depend on i's mental states in outcomes x and y may classify mental states solely in terms of their intrinsic properties, or also in terms of their relational properties. Robert Nozick's famous "experience machine" hypothetical is meant to suggest that well-being is not solely a matter of the intrinsic properties of mental states. Wayne Sumner, sensitive to this point, analyzes well-being in terms of "authentic happiness." As between two happiness states that are intrinsically identical, one may be authentic, e.g., based on true beliefs (a relational property), the other inauthentic.

i) and (*y*; *i*) depend on the type of mental states that *i* experiences in the two outcomes. Indeed, it draws a very strong link indeed between well-being and mental states. According to this theory, well-being *supervenes* on mental states. [66] On this theory, if *i*'s mental states in *x* and *y* are intrinsically identical, then *x* and *y* are equally good for *i*'s well-being. Yet the theory also makes the well-being ranking of (*x*; *i*) and (*y*; *i*) depend on *i*'s preferences as between *x* and *y*.

On the other hand, these two potential determinants of well-being – preference satisfaction and mental states -- are distinct. It is certainly possible for an account of well-being to make the ranking of (*x*; *i*) and (*y*; *i*) depend upon *i*'s mental states in the two outcomes that figure in these life histories, without appealing to *i*'s preferences between the outcomes. Such would be a view that appeals to "pleasure" and "pain" understood in strictly sensationalist terms, rather than as preferred or dispreferred mental states.

It is also certainly possible for an account of well-being to ground the ranking of (*x*; *i*) and (*y*; *i*) in *i*'s preferences between the outcomes, without drawing a strong link to *i*'s mental states in the two outcomes. A critical point, here, is that a preference is simply a choice-relevant ranking . The content of a preference –what it is about outcomes that determine an individual's preference ranking – can be any feature of outcomes (or at least any feature that the individual can have a causal impact upon[67]).Individual *i* can prefer (*x*; *i*) to (*y*; *i*) in virtue of differences between the two outcomes other than differences between his own mental states. Therefore, an account that defines well-being in terms of preference satisfaction (be it actual or fully informed preferences) without restricting the content of the preferences will *not* necessarily satisfy a mental-state supervenience requirement, and may at the extreme reach the verdict that whether a given individual *i* is better or worse off with (*x*; *i*) as compared to (*y*; *i*) does not depend at all on his mental states.[68]

---

[66]To say that well-being supervenes on certain properties is to say that, if individual *i* has the very same properties of that type in *x* and in *y*, then (*x*; *i*) and (*y*; *i*) are equally good for well-being. There must be some difference in the individual's properties, of the specified type, for there to be a well-being difference. In discussing *mental-state* supervenience, throughout this chapter, I mean supervenience on the intrinsic properties of mental states. An account of well-being incorporates a mental-state supervenience requirement if it says: If *i*'s mental states in *x* and *y* are identical in their intrinsic properties, (*x*; *i*) and (*y*; *i*) are equally good. Whether well-being supervenes on mental states in this sense has been a salient issue in the philosophical literature, as crystallized by Nozick's experience machine.

[67] It might be argued that an individual cannot prefer *x* to *y* if the two outcomes are identical in terms of all features that the individual can causally influence – because, if so, there is no possible choice situation where the ranking would rationally motivate the individual to choose one action rather than another. Even if this restriction on the content of preferences is accepted, individuals can prefer many features of outcome other than their own mental states.

[68]Whether such an account satisfies a mental-state supervenience requirement, for a particular outcome set and population, will depend on the preferences of those individuals regarding that outcome set. But clearly a preference-based view that doesn't restrict the content of preferences won't *necessarily* satisfy a mental-state supervenience requirement. Imagine that individual *i*'s actual preferences regarding outcomes, as well as his fully informed and rational preferences, are in part driven by features of outcomes other than the intrinsic properties of his own mental states. (For example, individual *i* might have a preference for being physically healthy, a preference

Just as the preferentialist theme within philosophical scholarship about well-being is echoed in economics, so too is the mentalistic theme. A burgeoning body of work within economics and psychology – the so-called literature on "happiness" or "subjective well-being" -- employs survey data to measure individuals' mental states, and then correlates the answers with other individual characteristics. The typical surveys employed by scholars working in this area ask respondents to quantify their "happiness" or "life satisfaction" on a numerical scale. Although the normative commitments of those involved in this literature are diverse, a number of leading practitioners do seem to adopt a substantially or even exclusively mentalistic concept of individual well-being (for example, the view that well-being is a matter of experiencing positive affects, avoiding negative affects, and experiencing a sense of satisfaction with life); and do seem to adopt the view that government should orient policy around producing individual well-being thus understood.

A third philosophical approach about well-being, which Parfit refers to as the "objective list" approach, is to furnish some list of "goods" that are seen as intrinsic constituents of well-being. A leading modern example of this approach is John Finnis, who argues that "life," "knowledge," "play," "aesthetic experience," "sociability" and "practical reasonableness" and "religion" are the basic forms of human good. Note that how fully an individual realizes the goods on Finnis' list is not merely a matter of what mental states that individual has. Further, whether an individual more or less fully attains these goods is not reducible to the satisfaction of his actual preferences. It is clearly *not* true that (x; i) is better than (y; i), in light of Finnis' list of goods, just in case i actually prefers outcome x to y. Finnis underscores the difference between realizing objective goods and preference-satisfaction in his discussion of the objective good of "knowledge."

> Nor can one validly infer the value of knowledge from the fact (if fact it be) that "all men desire to know." The *universality* of a desire is not a sufficient basis for inferring that the object of that desire is really desirable, objectively good. Nor is such a basis afforded by the fact that the desire or inclination manifests, or is part of, a *deep* structure shaping the human mind, or by the fact that the desire, or the structure, is *ineradicable*, or by the fact that in whole or part the desire is (or is not) *common* to all animals, or by the fact it is (or is not) *peculiar* to human beings ….

---

which he retains under conditions of full information and rationality. Thus he prefers x to y, because his physical body is healthier in outcome x, even though his mental states are intrinsically identical in the two outcomes. On an unrestricted preference-based view, (x; i) is better than (y; i).) At the extreme, individual i's actual and fully informed preferences may be largely or even exclusively driven by features of outcomes other than his own mental states. (For example, individual i is an artist who cares only about producing a work of beauty, regardless of how he feels about it.)

Note that, even in this extreme case, it is an overstatement to say that individual i's well-being does not depend on his mental states. After all, a preference is a mental state. On the view under consideration, the basis for (x; i) being better than (y; i) is that individual i has a preference ranking and x is better than y in terms of that ranking. The point, rather, is that it is possible for an individual to have a preference ranking of outcomes such that what determines whether x is higher ranked than y has nothing to do with the individual's mental states in x and y.

> ….It is obvious that a man who [possesses knowledge] *is* better off (other things being equal) than a man who [doesn't], not just in this particular case or that, but in all cases, as such, universally, and *whether I like it or not*. Knowledge is better than ignorance. Am I not compelled to admit it, willy nilly? It matters not that I may be feeling incurious myself. For the understanding affirmation [of the value of knowledge] is neither a reference to nor an expression of any desire or urge or inclination of mine. Nor is it merely a reference to (or implied presupposition of) any desires that my fellows happen to have. …. It is a rational judgment about a general form of human well-being, about the fulfillment of a human potentiality. As such, it has in its own way the peremptoriness of all other rational judgments. It constitutes a critique of my passing likes and dislikes.

Other contemporary philosophers who offer lists of well-being constituents that are "objective" in this sense – neither reducing to an individual's mental states, nor to the satisfaction of that individual's actual preferences – are Martha Nussbaum, George Sher, and James Griffin. Nussbaum argues that the "central human capabilities" are: life; bodily health; bodily integrity; the senses, imagination and thought; emotions; practical reason; affiliation; play; and control over one's environment. Sher endorses a list suggested by Parfit: "moral goodness, rational activity, the development of one's abilities, having children and being a good parent, knowledge, and the awareness of true beauty." Finally, Griffin's is: accomplishment; "the components of human existence" (roughly, autonomy and physical integrity); understanding; enjoyment; deep personal relations.

Although the philosophers I have just described *do* characterize an individual's well-being in terms that reduce it neither to the individual's mental states, nor to his actual preference satisfaction, it would be incorrect to say that mental states and preference-satisfaction are wholly absent from these accounts. First, although the accounts now under discussion certainly do not satisfy a mental-state supervenience requirement – each implies that (*x*; *i*) can be better for *i* than (*y*; *i*) even though the intrinsic properties of individual *i*'s mental states are identical in the two outcomes – all of these accounts clearly make the occurrence of different types of mental states *one* of the significant sources of well-being.

Second, objective-good accounts of well-being do not necessarily deny any link between well-being and preference satisfaction. George Sher analyzes objective goods as those items that advance "near-universal, near-unavoidable goals." A goal is a choice-connected attitude which is similar, if not identical, to a preference. Griffin offers a different kind of nexus between objective goods and preferences, as we'll discuss in greater detail below. On Griffin's view, a normal human being who recognizes that one of his life histories is objectively better than another will come to prefer it. Other objective good theorists, too, might be happy to claim this sort of motivational connection between recognizing an objective good and developing a preference for it.

Pulling all this together, we might say that the philosophers who have presented objective good accounts of well-being never reduce an individual's attainment of objective goods to his mental states, to his actual preference satisfaction, or even to the satisfaction of his idealized

preferences, at least if the idealizing conditions are not designed to screen out idiosyncratic preferences. Even those objective good philosophers who draw an explicit link between objective goods and preferences, such as Griffin and Sher, clearly deny that (*x*; *i*) is objectively better for *i* than (*y*; *i*) if *i* has a preference for the first outcome which is fully informed but idiosyncratic – not widely shared. On the other hand, the characterization of objective goods just offered allows for various *other* kinds of linkages between objective goods and preferences.

It is also worth clarifying the connection between objective goods and *human nature*. The idea of using human nature to specify a good human life goes back to Aristotle. The best developed contemporary account of this sort is Thomas Hurka's. Hurka identifies an individual's good with the development of those attributes which are "essential to humans and conditioned on their being living beings": having a living body, and having "theoretical and practical rationality," i.e., the capacity for knowledge and for forming and acting on goals.

This human-essence criterion yields an intuitively jarring account of objective goods. As Hurka freely admits, the theory "does not find intrinsic value in pleasure, not even pleasure in what is good, nor does it find intrinsic disvalue in pain." Nor does it value happiness or enjoyment. Finally, it values love and friendship, not as a source of emotional support, but because these are occasions for *teamwork*: for formulating and acting on goals with others. But, intuitively, family relations and deep friendships that are not particularly collaborative are (or at least can be) much more important for well-being than relationships with mere collaborators.[69]

However, it bears emphasis that the (implausible) strategy of analyzing well-being in terms of human nature is simply one *kind* of objective good approach. An account of well-being is characterizable as an objective-good account – I have suggested – if it does not reduce an individual's well-being to his mental states, or to the satisfaction of that individual's preferences (even idealized). An account might have these features *without* identifying objective goods as those goods that realize human nature. Indeed, the various prominent objective-good theorists I mentioned earlier – Finnis, Nussbaum, Sher, Griffin – do *not* specify objective goods along the lines that Hurka pursues. And the account of well-being that I will ultimately defend in this chapter – analyzing well-being in terms of individuals' fully informed, fully rational, convergent extended preferences – *is* indeed characterizable as an objective good account, but it certainly does not use human nature to identify the sources of well-being.

So much for the philosophical literature on objective goods. The proposition that well-being is a matter of objective goods is also reflected in much scholarship outside philosophy. Amartya Sen's theoretical work on "capabilities" and "functionings" seems to involve an objective-good account of well-being, in the sense just characterized. Sen's work, in turn, along

[69] Actually, Hurka's theory is not intended as a theory of well-being. He is careful to distinguish between an individual's good and individual well-being, and claims to be analyzing only the former. But the very fact that Hurka draws this distinction, and the particular list of goods he generates, suggests that using human nature to identify a list of the sources of human well-being is problematic.

with Martha Nussbaum's, has been the direct inspiration for a large body of recent empirical work in development economics and other areas, which seeks to measure how individuals are faring with respect to capabilities/functionings – as opposed to happiness/life-satisfaction or income (income being the traditional measure used by economists who equate well-being and preference satisfaction, as in cost-benefit analysis, the measurement of income inequality, or the measurement of income poverty). I think it fair to say that a kind of objective-good approach to well-being seems to be characteristic of much of this "capabilities"/"functionings" literature. So, too, is it characteristic of a different and somewhat older literature – the literature regarding "social indicators," in which various indices of the "quality of life" are constructed and empirically deployed. Robert Cummins systematically reviewed scholarship that employed multidimensional indices of "quality of life," and found that the different dimensions used in the scholarship generally could be placed within one of seven quality of life "domains": material well-being, health, productivity, intimacy, safety, community, emotional well-being. This list of quality of life domains looks like a somewhat more aggregated version of Nussbaum's list of "central human capabilities."

Why the Debate?

For purposes of this book, an account of well-being is an element within a *moral* decisional framework. I have adopted a reflective-equilibrium methodology for constructing this framework. In particular, then, an attractive account of well-being – for purposes of this book – is one that readers can endorse in reflective equilibrium.

But reaching a point of reflective equilibrium with respect to the nature of well-being is *difficult*. It is difficult to construct an account of well-being which satisfies various basic principles that seem intuitively attractive, and which is also consistent with intuitions about concrete cases. This difficulty is *evidenced* by the philosophical debates about well-being and, indeed, helps *explain* these debates – since many of the participants, implicitly or explicitly, are trying themselves to reach a point of reflective equilibrium.

One very plausible basic principle about well-being is that it has *critical* force. In other words, an individual can be mistaken about his own well-being. Although many economists may reject the principle, it is one which most philosophers of well-being are prepared to endorse. This is true even of philosophers who make preference-satisfaction a central element of their account. For example, Railton writes:

> [On one account, to] call something part of someone's intrinsic good [is] to say that he desires it for its own sake. This theory has many virtues: it is uncomplicated, nonpaternalistic, and epistemically as straightforward as the idea of desire ….
>
> Yet this theory is deeply unsatisfactory, since it seems incapable of capturing important elements of the critical and self-critical character of value judgments. On this theory one can, of course, criticize any particular current desire on the grounds that it ill fits with other, more numerous or more powerful current desires on one's part, or (if it is an instrumental desire) on the grounds that it is the result of a

> miscalculation with the information one has. But this hardly exhausts the range of assessment. Sometimes we wish to raise questions about the intrinsic desirability of the things that now are the main focus of our desires, even after any mistakes in calculation have been corrected. This appears to be a specific function of the vocabulary of goodness and badness, as distinct from the vocabulary of desire and aversion.

Note, however, that the idealizing conditions offered by Railton – and by other leading examples of philosophical preferentialists about well-being, such as Sidgwick, Rawls, and Brandt -- are *procedural*. By this term I mean conditions such as having good information; having preferences over outcomes which meet formal conditions such as being asymmetric or transitive; having coherence between first- and higher-order preferences; having preferences over choices which are linked to preferences over outcomes consistent with formal conditions, e.g., the conditions specified by expected utility theory; reasoning in a manner which is consistent with the norms of deductive logic; or having a certain kind of emotional state, e.g., a calm state suitable for deliberation.

Intuitively, there is a difference between such procedural conditions regarding idealized preferences and *historical* conditions (such as requiring that the holder of the preferences not have been the subject of certain kinds of parental or social influence); and both, in turn, are different from *substantive* conditions on preference formation, such as stipulating that a fully rational preference cannot be an intrinsic preference to eat a saucer of mud, to count blades of grass, or to be happy but only on Tuesday. Let us try to capture this difference by saying that procedural conditions for idealizing preferences depend on the occurent mental state of the holder of the preference and are content-neutral (i.e., are logically consistent with every possible ranking of outcomes); historical conditions depend on the past history of the holder of the preference and are content-neutral; substantive conditions are not content neutral.

A substantial number of philosophers suggest that defining well-being in terms of procedurally idealized preferences is insufficient to capture the critical force of well-being. This is true, in particular, of leading philosophers within the objective-good camp. For example, Griffin writes:

> [A] particularly irrational desire – say one planted deep when one was young – might well survive criticism by facts and logic, and its mere endurance is less than it takes for its fulfillment to make one better off. For instance, I might always wish to hog the limelight and have learned from long experience, perhaps even learned deeply from years of psychoanalysis, how this harms my life. But I might, none the less, still want to hog it. I might not react appropriately, or strongly enough, to what I have learned.

Richard Arneson argues that the following point is the "decisive objection" against informed desire accounts of well-being.

> [T]he essence of the informed-desire view is that what the process of becoming fully informed and critically reflecting causes one to desire for its own sake is good for one …. But nothing bars this casual process from generating outcomes in a way that does not intuitively confer any desirability on the resultant basic desires. It might simply be a brute psychological fact about me that if I were to become fully informed about grapes, this process would set off a chemical process in my brain that would lead me to

> crave counting blades of grass on courthouse lawns as my primary life aim. This would seem to be an oddity of my brain, not an indicator of my true well-being. If this were true of everyone, not just me, the same point would still hold. The informed-desire theories purport to establish that a certain causal process confers desirability; but the characterization of the causal process does not secure this result, and it does not seem that it could be altered to guarantee the right result.

Amartya Sen has famously argued that an account of well-being needs to be sensitive to the problem of adaptive preferences.

> A person who has had a life of misfortune, with very little opportunities, and rather little hope, may be more easily reconciled to deprivations than others reared in more fortunate and affluent circumstances. The metric of happiness may, therefore, distort the extent of deprivation, in a specific and biased way. The hopeless beggar, the precarious landless labourer, the dominated housewife, the hardened unemployed or the over-exhausted coolie may all take pleasures in small mercies, and manage to suppress intense suffering for the necessity of continuing survival, but it would be ethically deeply mistaken to attach a correspondingly small value to the loss of their well-being because of this survival strategy. The same problem arises with the other interpretation of utility, namely, desire-fulfillment, since the hopelessly deprived lack the courage to desire much, and their deprivations are muted and deadened in the scale of desire-fulfillment.

This observation, too, provides a critique of procedural idealization. An individual's history and socialization might be such that she has adapted to a life which seems to furnish relatively little well-being, and yet which she prefers – even with good information, reasoning calmly, having formally coherent preferences, and so forth.

In sum, the platitude that well-being has critical force pushes us away from an actual-preference account of well-being, towards an account that adds procedural idealizing conditions or even stronger idealizing conditions.

Unfortunately, this platitude is in some tension with a second one: that well-being has *motivational* force. If individual *i* is not, at some level, disposed to choose outcome *x* over outcome *y*, then it is counterintuitive to say that individual *i* is better off with life history (*x*; *i*) as opposed to (*y*; *i*). This principle should seem plausible, I suggest, independent of a more generic principle that moral considerations generally have motivational pull. Many metaethicists, so-called "internalists" about morality, adopt this more generic principle. But the platitude I am describing here involves "internalism" (of some sort) about *well-being*, which does not entail internalism about morality more generally – and indeed, I think, is intuitively more compelling than internalism about morality generally.[70] Imagine that I feel no inclination towards outcome *x* as opposed to *y*; it exerts no motivational pull on me, even when I am vividly informed about the two outcomes, think long and hard about them, and so forth. How, then, can *x* be possibly better than y *for me*?

[70] This is not to say that internalism about morality lacks appeal – just that internalism about well-being is even more compelling.

It is worth noting, at this point, that well-being has a relational character. Individual *i* realizes more well-being in outcome *x* than *y* iff outcome *x* is better *for individual i* than outcome *y*. What it means for one outcome to better *for* an individual than another outcome is not just a matter of the first outcome being better than the second in virtue of some attribute of that individual. For example, outcome *x* may be more beautiful than outcome *y*, in virtue of the fact that Jim is more beautiful in *x* than *y*, but this doesn't entail that *x* is better *for Jim* than *y*. (Jim's beauty may have no resonance for Jim: he may be indifferent to it, not have invested effort in producing it, etc.) Admittedly, what it means for an outcome to be better for an individual is elusive. But one aspect of the "good-for" relation (at least in the case of persons, as opposed to other well-being subjects) is the tie to motivation. Some outcome is not better for a person unless she is disposed to choose it (at some level).

An actual-preference account of well-being gives insufficient critical force to well-being, but has the great virtue of according with the motivational platitude. Philosophers who link well-being to *idealized* preferences presumably hope to satisfy the motivational platitude while also bolstering well-being's critical force. Railton stresses the motivational force of well-being when he writes:

> Is it true that all normative judgments must find an internal resonance in those to whom they are applied? While I do not find this thesis convincing as a claim about all species of normative assessment, it does seem to me to capture an important feature of the concept of intrinsic value to say that what is intrinsically valuable for a person must have a connection with what he would find in some degree compelling or attractive, at least if he were rational and aware. It would be an intolerably alienated conception of someone's good to imagine that it might fail in any such way to engage him.

It is trivially true that there is a link between *idealized* preferences and motivation under *idealized* conditions. A preference just *is* a choice disposition, and so the premise that I would prefer *x* to *y* under conditions *C* implies that, under conditions *C*, I would be disposed to choose *x* over *y*. Interestingly, however, an ideal-preference account may also have the upshot that well-being has wider motivational force – that it has force both under the idealizing conditions, and under less-ideal conditions that individuals in the actual world actually attain with some frequency. Railton suggests that individuals, under ordinary conditions, are normally motivated by knowledge of what they would want for themselves under the idealized conditions that he outlines: "[T]his notion of someone's good [which Railton has argued for] satisfies an appropriate internalist constraint: ….[T]he views we would have were we to become free of present defects in knowledge or rationality would induce an internal resonance in us *as we now are*." Brandt makes a similar claim: that for an individual under nonideal conditions to know that some of his desires would be lost under the ideal conditions of "cognitive psychotherapy" normally motivates the individual under nonideal conditions.

> By definition 'irrational' desires are one and all ones that the person would lose if he repeatedly reminded himself of known facts about himself or the world ….

> The proposal here is that awareness of the fact that one has irrational desires works in a way similar to awareness that one has incoherent beliefs or unjustified fears. One is made uncomfortable by the awareness and is motivated to remove its source. I am not offering any reason why this should be the case. I am asserting that, as a fact, people …. do dislike having to think that their desires are irrational in my sense ….

More generically, the claim that an individual, under ordinary conditions, is motivated by a belief about what he prefer under idealized conditions is a central aspect of Michael Smith's ideal-approval account of moral facts.

A third basic principle which has intuitive force, and which anyone constructing an account of well-being should aim to satisfy, might be termed the principle of non-remoteness. An individual's well-being must not be too remote from him. Parfit stresses this point in a well-known example.

> Suppose that I meet a stranger who has what is believed to be a fatal disease. My sympathy is aroused, and I strongly want this stranger to be cured. We never meet again. Later, unknown to me, this stranger is cured. On the [actual preference theory of well-being], this event is good for me, and make me life go better. This is not plausible. We should reject this theory.

Many other contemporary philosophers of well-being concur with Parfit than a simple preference-based account of well-being -- along the lines of (x; i) is better than (y; i) iff individual i prefers outcome x to outcome y -- is problematic because such an account allows events that are too remote from the holder of the preference to benefit him. For example, Darwall writes:

> There are many things I rationally take an interest in, such as the survival of the planet and the happiness of my children long after I am dead, that will make no contribution to my welfare. A person may have rational *interests* that go well beyond what is for her good or *in her interest*. A person's good – what benefits her or advances her welfare – is different from what is good from her point of view or standpoint. The latter is the perspective of what she herself cares about, whereas her own good is what is desirable from the perspective of someone (perhaps she herself) who cares for her.

Scanlon writes:

> [Desire theories of well-being] are open to serious objection. The most general view of this kind – it might be called the unrestricted actual-desire theory – holds that a person's well-being is measured by the degree to which all the person's actual desires are satisfied. Since one can have a desire about almost anything, this makes an implausibly broad range of considerations count as determinants of a person's well-being. Someone might have a desire about the chemical composition of some star, about whether blue was Napoleon's favorite color, or about whether Julius Ceasar was an honest man. But it would be odd to suggest that the well-being of a person who has such desires is affected by these facts themselves (as opposed to the pleasure he or she derives from having certain beliefs about it). The fact that some distant star is made up of the elements I would like it to be does not seem to make my life better (assuming that I am not an astronomer whose life work has been devoted to a theory that would be confirmed or refuted by this fact).

Arneson writes: "[N]ot all of an agent's desires bear on her well-being. I might listen to a televised plea for famine relief, and form the desire to aid distant starving strangers, without

myself thinking (and without its being plausible for anyone else to think) that the fulfillment of this desire would in any way make my life go better."

What exactly is the difficulty that "remoteness" poses for a simple actual-preference account of well-being? We return to the point that preferences, per se, are unrestricted in their content. The sheer fact that *i* prefers outcome *x* to outcome *y* imposes no constraint on the features that differentiate *x* and *y* (or none except a requirement that the differentiating features be subject to *i*'s causal influence.) Individual *i* might prefer *x* to *y* even though *i*'s mental states, and the condition of his physical body, are identical in their intrinsic characteristics in the two outcomes, as are the mental states and physical bodies of his friends and family. This is Parfit's stranger on the train case. A yet more powerful illustration of the remoteness objection involves posthumous preferences. Individual *i* can prefer *x* to *y* even though *x* and *y* differ only with respect to events that occur after his death.[71] (For example, *i*, who dies in the year 2000, prefers that the Grand Canyon never be despoiled; in outcome *x*, the Grand Canyon is never despoiled, in outcome *y* it is despoiled in the 24th century.)

A different aspect of the "remoteness" worry is that preferences, per se, can be grounded in a wide range of rationales. Individual *i* might prefer *x* to *y* for moral reasons, legal reasons, aesthetic reasons, because of social pressure, and so forth. Imagine that I am a member of a 5-member town council, trying to decide whether to spend funds to improve a school or a senior center. Neither I, nor anyone else on the council, have children in the school or relatives who use the center. I determine that, on balance, I have moral and legal reason to spend the funds on the school, because the state constitution requires an adequate public education, and because most children at the school are impoverished, while those who use the senior center are more affluent. Two other members of the council share my views, and the school funding is approved. My (morally and legally motivated) preference for the school has been satisfied. Am I better off than if the center had been funded?

Note, too, that idealizing preferences by adding more information, or by requiring that preferences be highly rational in a variety of procedural senses, does not cure the "remoteness" worry.[72]

But why exactly should we endorse the principle that an individual's well-being must not be too "remote" from him, and thus reject an account that ties well-being to actual or ideal preferences without restricting the content of preferences or their underlying rationales?

[71] Note that future events, including events that occur after an individual's death, are subject to an individual's causal influence, and thus can determine a preference ranking even if the content of a preference must be limited to features of outcomes that the individual can causally influence.

[72] Some procedural restrictions – for example, requiring that the preference be the result of a certain kind of attitude, such as self-sympathy – may cure the remoteness problem. But many standard procedural restrictions will not do so.

To begin, the "remoteness" principle is supported by our intuitions about particular cases (such as Parfit's stranger case and the others just described). Second, the principle is, intuitively, connected to the relational character of well-being.[73]

Finally, there is a serious tension between preference-based accounts of well-being which have not been structured to handle the remoteness objection and the possibility of non-self-interested reasoning and action. This is a point which Mark Overvold has stressed, using the example of self-sacrifice as a vivid illustration. I will illustrate Overvold's point using a simple actual preference account of well-being, which says that (x; i) is at least as good for individual i's well-being as (y; i) just in case i prefers outcome x to y. The problems I am about to illustrate generalize to *ideal-preference* accounts that have not been structured to handle the remoteness problem.

Consider the case in which the heroic soldier i prefers an outcome x in which he dies and his comrade is saved, as opposed to an outcome y in which he survives and his comrade dies. He chooses an action a which he knows will lead to x, or believes will do so with a high probability, as opposed to an action b which he knows will lead to y, or do so with a probability. (Action a is, say, leaping on a grenade which threatens the comrade; action b is not doing so.) The hero chooses action a, and outcome x occurs. Yet on the simple actual-preference account of well-being, the hero is better off in x – because he prefers x to y – and thus is not a hero at all. In short, this account makes it impossible to be a hero – that is, to engage in self-sacrifice, deliberately choosing an action which yields an outcome that is dramatically worse for the actor than the outcome of some other choice.

The most striking examples of self-sacrifice involve an actor who is wholly partial to someone else's well-being, and ignores his own interests entirely. But there is a tension between preference-based accounts which have not been restructured to handle the remoteness objection and *any* kind of non-self-interested reasoning and choice.[74] Consider moral reasoning. Moral reasoning is supposed to be impartial. I am supposed to give no greater weight to anyone else's interests than my own. Presumably it is a constitutive condition of genuinely impartial reasoning that is *possible* (although not necessary) for me to end up making a choice which is worse for my own interests. Note, therefore, how a simple actual-preference account of well-being precludes an individual from engaging in genuinely impartial reasoning. I might try to adopt an impartial perspective, and rank the outcomes in a given outcome set in an impartial manner; but, on this account of well-being, whatever ranking I arrive at will correspond exactly to how the outcomes are ranked in light of my own well-being. And if, after this exercise in attempted impartiality, I choose action a in light of my outcome ranking -- using the most attractive theory of how to rationally select choices in light of an outcome ranking, be it expected

[73]Intuitively, the problem with remote features of outcomes that individuals may prefer is that they don't make the outcomes better *for* the persons involved.

[74] Indeed Overvold clearly says as much.

utility theory or some other theory -- I will have chosen an action which is (ex ante) best for my well-being.

A major difference between philosophical writing about well-being, and welfare economics, is that philosophers have been much more sensitive to the remoteness worry than economists. Scholarship on social welfare functions, cost-benefit analysis, or other aspects of welfare economics rarely, if ever, builds in a condition on the content of preferences or their underlying rationale which is explicitly designed to distinguish between preferences that are welfare-enhancing and preferences that are not. [75]

How should the remoteness worry be handled, in constructing an account of well-being? One possibility is via a supervenience requirement, which says that (*x*; *i*) and (*y*; *i*) can differ for *i*'s well-being only if the two outcomes differ with respect to certain sorts of facts. This observation helps shed light on why theories of well-being that draw a link to mental states have some real attractiveness. What could be more "proximate" to a person than his mental states? Consider a theory which says that (*x*; *i*) is better for *i* than (*y*; *i*) iff *i*'s mental states in the two outcomes are different, in terms of their intrinsic properties, in some stipulated way --- meaning for example that the mental states in *x* feel better to *i*, or that *i* prefers the mental states in *x* to *y* in virtue of the intrinsic differences between these mental states, or would prefer them under ideal conditions. Seemingly, whatever other worries one might have about such an account, "remoteness" is not one of them.

But the theory now on the table goes too far. In its zeal to circumscribe the features of outcomes that can affect the ranking of an individual's life-history, it leaves out those non-mentalistic features that, intuitively, are still welfare relevant. This is the nub of Robert Nozick's famous "experience machine" objection to mental state theories.

> Suppose that there were an experience machine that could give you any experience you desired. Superduper neuropsychologists could stimulate your brain so that you would think and feel you were writing a great novel, or making a friend, or reading an interesting book. All the time you would be floating in a tank, with electrodes attached to your brain. Should you plug into this machine for life, preprogramming your life's experiences? If you are worried about missing out on desirable experiences,

[75] The remoteness problem is sometimes finessed, sub rosa, in the literature by making each individual's preferences for outcomes depend solely on individual attributes that are clearly non-remote. For example, in the canonical set-up for the fundamental welfare theorems, each individual's preferences are a function solely of his consumption of different goods; and, in much of the optimal tax literature, an individual's utiltiy function depends on his own consumption and leisure. However, there is rarely systematic discussion of the proposition that the content of preferences needs to be thus limited so as to deal with the remoteness problem. And there are certainly other contexts, in welfare economics, where the content of preferences is *not* thus limited. A very important example is "existence values" in cost-benefit analysis – where WTP/WTA values are elicited for environmental characteristics that may well be spatially or even temporally distant from the individuals whose WTP/WTA values are being elicited (for example, the preservation of a remote ecosystem or the saving of an endangered species with which the individual has never interacted).

> we can suppose that business enterprises have researched thoroughly the lives of many others. You can pick and choose from their large library or smorgasbord of such experiences…. Would you plug in?

Nozick concludes: "We learn that something matters to us in addition to experience [mental states] by imagining an experience machine and then realizing that we would not use it."

Shelly Kagan and Mark Overvold have suggested other kinds of supervenience requirements to handle the remoteness objection. Kagan tentatively suggests that (*x*; *i*) is better for *i* than (*y*; *i*) only if the two outcomes differ either in terms of the intrinsic properties of *i*'s mental states, or in terms of the intrinsic properties of his physical body. Overvold argues that well-being consists in the satisfaction of "self-interested" preferences, defined as follow:

> [T]he only desires and aversions that are logically relevant to the determination of an individual's self-interest are those in which (1) it is logically necessary that the individual exist at t for the object of one's desire or aversion to obtain at t, and (2) the reason for this desire is due to one's essential involvement in the state of affairs.

Translated into a condition on the ranking of outcome sets, Overvold's account says something like the following: Individual *i*'s ranking of an outcome set is self-interested if (1) for any pair of outcomes (*x*, *y*) such that *i* is not indifferent between the two, there is some difference between them that occurs when *i* is alive; and (2) individual *i*'s ranking of the pair of "nearest possible" counterpart outcomes in which *i* does not exist, call them (*x*+, *y*+), is such that *i* is indifferent between (*x*+, *y*+).

Both of these proposals, like the mental-state supervenience proposal, are counterintuitive in some cases. Kagan's proposal is counterintuitive because, seemingly, my well-being can be affected by events outside both my mind and my body. (A standard example is where someone's family, friends, or colleagues do or say things without her being aware of them. Imagine that Sheila is betrayed by her spouse, but believes that he is faithful, and never learns otherwise. Isn't the fact of the betrayal itself something that makes Sheila's life worse, independent of her learning of the betrayal?) Overvold's account is counterintuitive, as well as Kagan's and the mental state account, because, seemingly, I can have non-self-interested preferences regarding what happens to my mental states or physical body. Imagine the penitent wrongdoer who develops a moral preference that he suffer pain and anguish in punishment for the wrongdoing. Surely *that* preference is not self-interested -- but note that it fails to be screened out by any of the proposed responses to the remoteness problem that we have thus far considered: Overvold's proposal, Kagan's, and a mental-state supervenience requirement.

A different approach to handling the remoteness objection is to couple an idealized-preference account of well-being with some attitudinal restriction. There is a long philosophical tradition, going back to Adam Smith, which draws a link between well-being and *sympathy*. A important recent example is Stephen Darwall, who suggests (1) that sympathy is indeed a "natural" attitude, in the sense of not essentially involving beliefs about well-being or other

normative beliefs; (2) that the sympathetic person can be partial in his sympathy and, at the extreme, exclusively sympathetic towards one particular person; and (3) that a person can be self-sympathetic. Pulling together these ideas, we might say that (*x*; *i*) is at least as good for *i* as (*y*; *i*) iff *i* prefers the first outcome under conditions of full information and full rationality, and under conditions where *i* is exclusively self-sympathetic.

A final possibility is to handle the remoteness objection by coupling an ideal-preference account of well-being with a *value-laden* account of the formation of those preferences. We might say that (*x*; *i*) is at least as good for *i* as (*y*; *i*) iff *i*, under conditions of full information and rationality, *judges x to be better for his well-being* and thereby comes to prefer *x*. Presumably *i*, in reaching this idealized judgment about his own well-being, is insensitive to remote features of *x* and *y* (for example, features that occur after his death.) The viability of *this* strategy for handling remoteness hinges on the viability of a value-laden account of the preferences that undergird welfare – a topic to which we now turn.

Metaethical Disputes

The philosophical literature on well-being is not merely characterized by first-order disagreements about which account is most attractive. A further source of disagreement is metaethical.

In my brief overview of metaethics, in Chapter 1, I discussed ideal-approval accounts of moral facts, and mentioned that so-called "secondary quality" accounts are a particular variant of ideal-approval accounts.

Remember that metaethical views can be cognitivist or noncognitivist. Cognitivists see moral "assertions" as genuine assertions of moral facts; believe that moral facts exist; and therefore believe that moral assertions can be true or false, no less so than paradigmatic factual assertions, e.g., statements about the physical world. Ideal approval accounts are one family of cognitivist accounts. Such accounts have the common feature that they analyze moral facts in terms of the approvals of some individual or group of individuals, reasoning under conditions that are idealized in some way. The statement "Item *m* is morally good," with *m* an action, an outcome, or some other item, means something like: "Individual(s) reasoning under conditions *C* would approve *m*."

An important distinction *within* the family of ideal-approval accounts is the distinction between "secondary quality" accounts (sometimes known as "sensibility" accounts) and other views. What is this distinction? It concerns whether the idealized reasoning specified by conditions *C* is supposed to be *value-laden* or *value-free*.

"Secondary quality"/sensibility accounts see these conditions as *value-laden*. More precisely, such accounts argue that the idealized individuals whose approvals undergird moral

facts are engaged in a reasoning process that involves moral judgments or perceptions. "Item *m* is morally good" means something like "Individuals under conditions C would perceive or judge *m* to be morally good, and thus would approve *m*." The proponents of secondary-quality accounts of moral facts often draw an analogy to facts about *colors*. Plausibly, the statement "Item *m* is red" means "Item *m* would look red to individuals under conditions *C'*. They would perceive *m* to be red." Similarly, it is suggested, the moral rightness or goodness of some item means that individuals under suitable conditions would see or judge it to be morally right or good. For example, David Wiggins writes:

> In so far as Hume ever came anywhere near to suggesting a semantical account of "x is good/right/beautiful" … it may seem that the best proposal implicit in his theory of valuation is that this sentence says that *x* is the kind of thing to arouse a certain sentiment of approbation. …
>
> …What after all is a sentiment of approbation? … Surely a sentiment of approbation cannot be identified except by its association with the thought or feeling that *x* is good (or right or beautiful) and with the various considerations in which that thought can be grounded. ….
>
> In all these matters, an analogy with colour is suggestive. "x is red if and only if x is such as to give, under certain conditions specifiable as normal, a certain visual impression" naturally raises the question "which visual impression?" And that question attracts the answer "an impression as of seeing something red," which reintroduces *red*. But this finding of circularity scarcely amounts to proof that we can after all appeal to something beyond visual impressions to determine colour authoritatively.

A different sort of ideal-approval view stipulates that the idealized reasoning giving rise to moral facts is *value-free* – more precisely, that reasoning under conditions *C* does *not* involve moral judgments or perceptions. For example, Ronald Milo reduces moral facts to what would be chosen in an ideal contracting scenario and, in so doing, very clearly articulates a *value-free* variant of an ideal-approval construal of moral facts.

> What the moral facts are – for example, which acts are wrong – is determined by which principles would be chosen by the hypothetical agents … It must be noted, however, that the moral principles chosen by the hypothetical contractors are viewed by them as *action guides*, not as *truth claims.*. The agents of construction are not to be conceived of as trying to reach an agreement on which moral principles are true, since apart from their agreement there are no antecedently given moral truths for them to discover. Rather, they determine through their choices which moral principles are true ….
>
> … [The contractors' reasoning process] will include certain desires that all (normal) human beings can be presumed to share…. But it will not include any moral beliefs about what is right or wrong in our interactions with others or about which traits of character are other-regarding virtues. The hypothetical contractors are not presumed to have any such beliefs.

Critics of a secondary-quality account of moral facts argue that such accounts are viciously circular or, at the very least, unilluminating. Proponents of such accounts argue that there is no vicious circularity; that it is impossible to provide an accurate characterization of good moral reasoning without allowing moral concepts, beliefs, etc., to figure in this reasoning; and that the hope to ground moral facts in idealized value-free reasoning is therefore chimerical.

How do these metaethical disputes relate to controversies about well-being? The value-free variant of an ideal-approval account of moral facts is naturally paired with an account of well-being in terms of idealized and *value-free* preferences – value free in the sense that the preferences are not themselves grounded in judgments or perceptions of well-being. If one worries that the secondary-quality analysis of moral facts is viciously circular or unilluminating, one should also worry that the following account of well-being is viciously circular or unilluminating: "One life history is better for the subject's well-being than a second iff the subject, under stipulated conditions, would judge or perceive the first to be better for well-being."

Indeed, Richard Brandt – who suggests an analysis of well-being in terms of preferences that are fully informed and fully rational in the sense of surviving "cognitive psychotherapy" -- characterizes that process as follows:

> This whole process of confronting desires with relevant information, by repeatedly representing it, in an ideally vivid way, and at an appropriate time, I call *cognitive psychotherapy*. I call it so because the process relies simply upon reflection on available information, without influence by prestige of someone, use of evaluative language, extrinsic reward or punishment, or use of artificially induced feeling-states like relaxation. It is *value-free reflection*.

Conversely, if one is drawn to a secondary-quality view of moral facts, then an analysis of well-being in terms of idealized value-free preferences will also, naturally, be seen to be wrongheaded. If moral facts are not reducible to ideal reasoning shorn of moral concepts, then why would well-being facts be reducible to ideal reasoning shorn of well-being concepts? Indeed, the central thrust of James Griffin's critique of traditional preferentialist accounts of welfare is that the preferences giving rise to well-being are value laden. As Griffin argues (focusing on the good of accomplishment):

> The taste model assumes that we can isolate valued objects in purely natural terms and then, independently, react to them with approval or disapproval. But can we? Prudential deliberation about accomplishment is not a case of first perceiving facts neutrally and then desire's entering and happening to fix on one object or other. The act of isolating the objects we value is far from neutral. We bring what I am calling "accomplishment" into focus only by resorting to such terms as "giving life weight or point," and such language already organizes our experience by selecting what we see favourably. Desire is not left free to happen to fix on one object or another; its direction is already fixed in, and manifested by, what we see favourably.

And he continues:

> [D]esires are not independent of the recognition of the good. The very few desires of which this is not true – say, some baffling urge left by hypnotic suggestion – are only vestigial desires …. [I]f we were beset by mere urges, coming from we know not where and we know not why, inclining us this way and that, we should see them as something to resist, to rid ourselves of as much as we could.

It should be stressed that Griffin's account of well-being retains a link to preferences. His assertion that "[d]esire is not blind," i.e., that well-formed preferences for different lives are the result of using normative concepts like "the good life" or "well-being," is coupled with an assertion that "reason is not inert," i.e., that seeing some life as good brings motivation in its train, at least for normal humans.

> Values are … what one would want if one properly appreciated the object of desire. …[A]s we have seen, this account shifts importance away from the mere occurrence of desire on to the nature of its object. Desire is left playing very little role, even *many* people's desires. Still, desire reappears in another place. To recognize the nature of the relevant object is to see it under some desirability characterization, such as "accomplishment" or "enjoyment." These desirability characterizations give reasons for action, and those reasons in turn mesh with characteristic human motivation.

Griffin's account of well-being is therefore a kind of full-information preferentialist account, with the critical proviso that the idealized conditions giving rise to preferences involve judgments or perceptions of well-being. Once more, there is an isomorphism to secondary-quality accounts of moral facts. The secondary-quality view, after all, is a kind of ideal-*approval* account. Seeing item *m* to be morally good produces an approval of *m*. Similarly, on Griffin's view of well-being, the fact that (*x*; *i*) is better for *i*'s well-being than (*y*; *i*) can be more or less analyzed as: "If individual *i* were characterized by normal human desires and concepts and were to reason about outcomes *x* and *y* with good information, he would perceive or judge the first outcome to be better for his well-being and thus come to prefer it."

Interpersonal Comparisons

The elements of well-being, and its metaethical basis, are topics of much philosophical debate, as we have just seen. By contrast, the possibility of interpersonal comparisons seems relatively uncontroversial. The topic has received relatively little sustained philosophical attention, but it seems that most philosophers – even those who make preference-satisfaction a central element of well-being – accept the possibility of comparing well-being levels and differences across persons.

Indeed, as I shall review in a moment, the prima facie case for interpersonal comparisons is a strong one. There a variety of different considerations, both intuitive and more systematic, that count in favor of both level and difference comparability.

Philosophers are not alone in acknowledging interpersonal comparability. As discussed in Chapter 2, the possibility of interpersonal comparisons is also accepted by some economists – in particular, economists who work with social welfare functions. However, there is an older tradition in economics that is skeptical about interpersonal comparisons. The so-called "ordinalist revolution" of the 1930s and 1940s was an attempt to orient positive and normative

economics around individual preferences, and without making interpersonal comparisons. The development of the Kaldor-Hicks criterion was one central element of this "ordinalist" school.

The failures of the Kaldor-Hicks criterion soon came to light, and indeed were a major impetus for the development, beginning in the 1970s, of the idea of a social welfare function using interpersonally comparable utilities as its inputs. But this approach is far from universally accepted by economists.

Why were the "ordinalists" skeptical about interpersonal comparisons? And why does their skepticism continue to resonate among many contemporary economists?

One source of the ordinalists' aversion to interpersonal comparisons , seemingly, was a concern that economics should be suitably "scientific." [**Discuss and criticize.]**

But there is a different and more creditable basis for skepticism about interpersonal comparisons: namely, that individual *preferences* provide no evident basis for such comparisons. For anyone who adopts a view of well-being that makes preferences a central element, there is indeed a genuine intellectual puzzle about the possibility of interpersonal comparability.

Consider a person's ranking of outcomes. Such a ranking immediately allows us to make sense of *intra*personal comparisons. We can say something like the following: life-history (x; i) is at least as good for well-being as life-history (y; i) iff the subject of both life-histories, i, weakly prefers outcome x to outcome y under suitable conditions (for example, when i is well-informed, rational, and self-interested).[76] But how on earth are we to make sense of an *inter*personal ranking of life-histories, within the confines of a preference based view? The sheer fact of having a preference over outcomes seems to be the very same thing as – or at least closely connected to -- a ranking of one's own life-histories; but no one, in ranking outcomes, produces a ranking of life-histories involving different individuals.

Harsanyi seeks to answer this challenge. In this section, I first review the prima facie case for interpersonal comparability, then discuss how Harsanyi employs the idea of an *extended preference* in his attempt to make sense of interpersonal comparisons. This discussion of Harsanyi's views, together with the review of the philosophical literature about well-being that has already been undertaken, will provide the underpinnings for the account of well-being that I shall present in the next section.

The Generic Case for Interpersonal Comparisons

In discussing interpersonal comparability, it is important to keep in the mind the distinction – standard within the social choice literature – between interpersonal comparisons of

[76] To be clear, the account of well-being that I ultimately present does not say exactly this. Rather, it says that (x; i) is at least as good as (y; i) iff individual i, under conditions of full information and rationality and self-interest, weakly prefers outcome x to y; *and* every other individual, under conditions of full information and rationality and under a condition of being i-interested, prefers outcome x to y.

well-being *levels* and interpersonal comparisons of well-being *differences*. Within my framework, that distinction has been framed as follows. An outcome set **O** and group of *N* individuals gives rise to a set **H** of life-histories. An account of well-being produces a quasiordering of **H**, denoted $\succcurlyeq^{WB}$: a ranking of life-histories. This account allows for *interpersonal comparisons of well-being levels* iff there is a least one pair of life-histories involving different subjects, such that the first life-history is ranked at least as good for well-being as the second. An account of well-being may also produce a difference quasiordering, $\succcurlyeq^{D}$, which is a ranking of *pairs* of life-histories. The relation $[(x; i), (y; j)] \succcurlyeq^{D} [(z; k) , (w; l)]$ means that the difference between the well-being of (*x*; *i*) and (*y*; *j*) is at least as large as the difference between the well-being of (*z*; *k*) and (*w*; *l*). An account of well-being allows for interpersonal *comparisons of well-being differences* iff it includes a difference quasiordering $\succcurlyeq^{D}$, and there is at least one such relation between pairs of life-histories where the four subjects are not identical.

Remember, too, my definition of the SWF approach to decisionmaking. Such an approach involves an account of well-being that allows for at least some interpersonal comparisons (of levels or differences). Moreover, it supplements the basic elements of the generic welfarist framework (outcome set, population, life-history set, and account of well-being) with a set **U** of utility functions. **U** represents the ranking of life-histories and difference set forth by the account. The SWF approach then ranks outcomes as a function of the utility vectors corresponding to the outcomes.

The "ordinalist" tradition within economics is skeptical of both interpersonal level comparisons and interpersonal difference comparisons. However, an account of well-being that declines to make either sort of comparison is problematic in various ways.

To begin, such an account is counterintuitive. It is easy, indeed trivial, to describe specific cases in which, intuitively, one person is at a higher well-being level than another. (Imagine a case in which one individual has a low income, bad health, suffers terrible pain, is socially stigmatized, and has no friends, while another has a high income, excellent health, feels great, and has a high social status and lots of friends.) It is also easy, indeed trivial, to describe specific cases in which the change in someone's well-being is greater than the change in someone else's well-being – a kind of interpersonal difference comparisons. (Imagine that individual *i*'s attributes in *x* and *y* are identical, except that *i* has slightly more income in *y*. By contrast, *j* has much more income in *x* than *y*, is in much better health, feels happier, lives many more years, and so forth. Then, intuitively, the difference between (*x*; *j*) and (*y*; *j*) is greater than the difference between (*y*; *i*) and (*x*; *i*).

More systematically, we saw in Chapter 2 that policy evaluation frameworks which decline to make interpersonal comparisons are problematic. Either these frameworks fail to meet the minimal welfarist standard of producing a Pareto-respecting quasiordering of outcome sets; or they do so, but in a manner that is less attractive than the SWF approach.

Why does the SWF approach itself require interpersonal comparisons? Why not structure moral decisionmaking by employing an account of well-being that makes only *intrapersonal* comparisons; by representing *that* account via a set **U**; by using **U** to transform outcomes into utility vectors; and by ranking outcomes as a function of their corresponding vectors? Remember the answer I provided in Chapter 2: the SWF framework without any interpersonal comparisons (of levels or differences) threatens to collapse to the Pareto quasiordering.

Finally, it bears note that rejecting interpersonal comparisons renders otiose many substantive debates within moral theory that philosophers have undertaken for many years. Much of moral philosophy since Bentham has involved a debate about utilitarianism, together with related debates to which this discussion has given rise (for example, the debate between welfarists and nonwelfarists, or consequentialists and deontologists). But utilitarianism involves interpersonal comparisons, as do the non-utilitarian variants of welfarism that have been discussed by philosophers and that correspond to a range of different SWFs reviewed in Chapter 4. Further, and notably, the philosophical critics of these various welfarist views as well as their proponents have generally accepted the possibility of interpersonal level and difference comparisons. The criticism has been substantive: that utilitarianism doesn't take seriously the difference between persons, that welfarist views ignore agent-relative constraints or the moral relevance of responsibility, and so forth. The critics have generally *not* claimed that utilitarianism and other welfarist views are just unintelligible, and can be rejected out of hand, because their presupposition of interpersonal comparability is false.

What about an account of well-being that allows for interpersonal level comparisons but not difference comparisons? Difference comparisons introduce an extra element of complexity into a decisionmaking framework. Moreover, as we'll see, how to construct utility functions that represent difference comparisons has been a particularly contentious issue among economists who accept interpersonal comparability. Why not avoid this controversy, and employ an SWF paired with an account of well-being that makes interpersonal level comparisons but does not make difference comparisons (or at least does not make interpersonal difference comparison)?

Here, we need to introduce some additional terminology. Let us say that interpersonal difference comparisons are "required" for a particular SWF if the SWF collapses to the Pareto quasiordering without difference comparisons.[77] A weaker idea is that interpersonal difference comparisons are "relevant" to a SWF. What "relevance" means is the following: If **U** contains some utility function $u(.)$, and we add a new function $u^*(.)$ to **U** which implies the very same

[77]If a moral decision procedure incorporates an account of well-being that makes interpersonal level but not difference comparisons, and represents the well-being quasiordering of life-histories via a set **U**, then for any such set **U**, supplementing **U** with every increasing transformation of every member of **U** represents the ordering of life-histories equally well. (If $u(.)$ is a member of **U**, to say that $v(.)$ is an "increasing transformation" of $u(.)$ means: $v(x; i) \geq v(y; j)$ iff $u(x; i) \geq u(y; j)$ for all life histories.) When I say that a given SWF "collapses to the Pareto quasiordering without difference comparisons," what I mean is: for any **U**, if that set is supplemented with every increasing transformation of every member, becoming **U*** , the SWF applied to **U*** is the Pareto quasiordering.

ranking of well-being levels as $u(.)$ but a different ranking of well-being differences, then the SWF's ranking of outcomes may change once **U** is supplemented by adding $u^*(.)$.[78]

There certainly are SWFs that do not require interpersonal difference comparisons. The leximin SWF is the leading example. Indeed, information about well-being differences is not even *relevant* to the leximin SWF.

However, the leximin SWF has substantial difficulties (see Chapter 4), and interpersonal difference comparisons *are* relevant to all the plausible competitors to the leximin SWF discussed in Chapter 4 – including standard functional forms such as the utilitarian SWF, a continuous prioritarian SWF, the rank-weighted SWF, and others. Why information about well-being differences is not relevant to the leximin SWF, but *is* relevant to these competing SWFs, is illustrated in the following table.

[Table]

Can we make a stronger claim? Is it true that interpersonal difference comparisons are *required* for plausible SWFs other than the leximin SWF – that these SWFs will "collapse" to the Pareto quasiordering absent such information? This does not seem to be strictly true; although it may be true that they "come close" to collapsing.[79] In any event, it is clearly true that information about interpersonal well-being differences is *relevant* to a wide range of SWFs other than the leximin SWF – and that is sufficient motivation to attempt to construct an account of well-being that encompasses such comparisons as well as interpersonal comparisons of well-being levels.

Moreover, as already mentioned, it is easy to construct cases in which, intuitively, one person's well-being change is greater than another's. This also cuts against the plausibility of a well-being account that solely makes interpersonal level comparisons.

---

[78] More precisely, interpersonal difference comparisons are relevant to a given SWF if: (1) there exists some **U**, such that if we supplement **U** with $v(.)$, where $v(.)$ is an increasing transformation of some member of **U**, the resulting set **U*** implies a quasiordering of well-being differences which is not the same as that implied by **U**; and (2) the SWF, applied to **U***, yields a different quasiordering of outcomes than when applied to **U**.

[79] For the sake of illustration, consider the utilitarian SWF. Imagine that we start with a set **U**. Assume that outcomes x and y are Pareto-noncomparable and that the utilitarian SWF, applied to **U**, ranks x as greater than or equal to y. Because the two are Pareto noncomparable, there exists some $u(.)$ in **U** and some individual i, such that $u(y; i) > u(x; i)$. Imagine, further, that, according to $u(.)$, there is no life-history in x or y, other than $(y; i)$, with the utility level $u(y; i)$. And there is no life-history in x or y, other than $(x ; i)$, with the utility level $u(x; i)$. Then presumably there is some v(.) which is an increasing transformation of $u(.)$, and which "stretches" the difference between $(x; i)$ and $(y; i)$, and "shrinks" the difference between $(x; j)$ and $(y; j)$ for all $j \neq i$, such that utilitarian SWF, applied to $v(.)$, says that y is better than x.

However, this "collapsing" argument does not work if some individuals in outcomes x and y have the same well-being levels. For example, imagine that there are two outcomes and two individuals, Abe and April. **U** consists of a single $u(.)$, such that $u(x; \text{Abe}) = 3$, $u(x; \text{April}) = 7$, $u(y; \text{Abe}) = 7$, $u(y; \text{April}) = 3$. Note that x and y are Pareto noncomparable. The utilitarian SWF, applied to **U**, ranks the two outcomes as equally good (not incomparable). Any increasing transformation of $u(.)$ will also rank them as equally good.

Finally, as we shall see in Chapter 4, a critical device for evaluating different kinds of SWFs is the *Pigou-Dalton principle* in terms of well-being. Whether an SWF satisfies this principle, either in general or below a threshold, is a critical question in the "reflective equilibrium" process of sorting between different kinds of SWFs. However, the Pigou-Dalton principle in terms of well-being presupposes both interpersonal level and difference comparisons. What is says is: If individual *i* is worse off than individual *j* in outcome *x*, and there is a pure transfer of well-being from *j* to *i* – such that individual *i*'s well-being is increased by a certain amount, with individual *j*'s well-being reduced by the very same amount, with no one else affected – and this transfer does not cause the individuals to "switch ranks,"[80] then the transfer is a moral improvement. Note, here, that characterizing an outcome as one where some individual is better off than another involves an interpersonal level comparison. And characterizing the change in some individual's well-being as being the very same amount as the change in another individual's well-being involve an interpersonal difference comparison.

Harsanyi's "extended preference" solution to the puzzle of interpersonal comparisons

Thus the case for interpersonal comparisons, both level and difference comparisons, is strong. John Harsanyi proposes to explicate comparisons of both kinds, and to represent such comparisons via utility numbers. The core of Harsanyi's proposal is the idea of an *extended preference* – a preference for life-histories or lotteries over life-histories -- and the use of expected utility (EU) theory to represent individuals' extended preferences.

This section presents these basic ideas, showing why they indeed hold promise as a basis for interpersonal comparisons. However, the section also outlines various problematic features of Harsanyi's views. Although Harsanyi is owed much credit for elaborating the idea of extended preferences and marrying it with EU theory, substantial work will be needed to incorporate these ideas into an attractive account of well-being -- the account that I will present in the next section.

Harsanyi imagines that there is a population of *N* individuals. A given member of the population is morally ranking various outcomes.

> Society consists of *n* individuals …. Suppose that individual [*k*][81] wants to make a *moral value judgment*. This will always involve comparing two or more social situations concerning their relative merits from a moral point of view. These social situations may be alternative patterns of social behavior …, alternative institutional frameworks, alternative governmental policies, alternative patterns of income distributions, and so forth. Mathematically, any social situation can be regarded as a *vector* listing the economic, social, biological, and other variables that will affect the well-being of the individuals making up the society. Different social situations will be called *A, B,* …

What Harsanyi calls a "social situation" is what I call an "outcome."

[80] It leaves individual *i* no better off than *j*.

[81] Harsanyi refers to the individual as "*i*," but I will generally refer to the spectator as "*k*," and so have changed his notation to be consistent with my own.

Harsanyi further supposes that individual *k* morally ranks outcomes by ranking life-histories and lotteries over life-histories – more specifically, by seeing each outcome as an equiprobability lottery over its component life-histories.

> Now if individual [*k*] wants to make a moral value judgment about the merits of alternative social situations *A, B*, …, he must make a serious attempt not to assess these social situations simply in terms of his own personal preferences and personal interests but rather in terms of some impartial and impersonal criteria. …
>
> Individual [*k*'s] choice among alternative social situations would certainly satisfy this requirement of impartiality and impersonality, if he simply *did not know in advance* what his own social position would be in each social situation – so that he would not know whether he himself would be a rich man or a poor man, a motorist or a pedestrian, a teacher or a student, a member of one social group or a member of another social group, and so forth. More specifically, this requirement would be satisfied if he thought that he would have an *equal probability* of being *put in the place* of any one among the *n* individual members of society.
>
> ….
>
> Thus [individual *k* will engage in] some process of *imaginative empathy*, i.e., by imagining himself to be *put in the place* of individual *j* in social situation *A*.
>
> This must obviously involve his imagining himself to be placed in individual *j*'s *objective position*, i.e., to be placed in the objective conditions (e.g., income, wealth, consumption level, state of health, social position) that *j* would face in social situation *A*. But it must also involve assessing these objective conditions in terms of *j*'s own *subjective attitudes* and *personal preferences* …..

Harsanyi uses the symbol "A_i" to denote "individual *i*'s personal position in social situation *A* (i.e., the objective conditions that would face individual *i* in social situation *A*)." This is what I term a life-history. A_i is a pairing of an individual with a "social situation" (outcome); and that is exactly how I have defined a life-history. And he refers to an individual's ranking of life-histories or lotteries over life-histories as her *extended preference* ranking. [82]

Assume that this idea of an "extended preference" is a coherent one. If so, it will be useful to have a term to refer to the person who is ranking life histories, and to distinguish that person from the individuals whose life-histories are being ranked. As previously, I will use the term "subject" to mean the latter individuals. I will use the term "spectator" to mean the former individual. These are my own terms, not Harsanyi's, but they help to explicate his ideas. If individual *k* possesses an extended preference ranking of life-histories, such as (*x*; *i*) and (*y*; *j*), then individual *k* is a "spectator"; individual *i* is the subject of life-history (*x*; *i*); individual *j* is

[82] Strictly, Harsanyi pairs each life-history with the subject's preferences. He terms this pairing an "extended alternative." An "extended preference," in turn, is a ranking of extended alternatives. My conception is more general: an extended preference ranks life-histories. The *description* of outcomes might include information about the subject's preferences. Or we might stipulate that everyone knows what everyone else's preferences are. But my conception of an extended preference also makes it possible that someone can possess an extended preference over other persons' life-histories without knowing what the subjects' preferences are. (Remember that an outcome, on my view, is a simplified possible world, which leaves out much information about subjects and background facts).

the subject of life-history (y; j). The term "spectator" does not denote some figure who is external to the population of N individuals, nor is there a single spectator. Rather (assuming the idea of an "extended preference" is coherent), *each* individual in the population of N possesses his own extended-preference ranking of the set of life-histories. Each individual in the population can function as a spectator. This is what Harsanyi supposes – and what my own account will suppose as well.

The final, critical, element of Harsanyi's analysis involves EU theory. EU theory, in the full form pioneered by Leonard Savage, provides a theory of rational choice under uncertainty, which has the following basic structure. Imagine that the decisionmaker is ranking an action set, in light of some outcome set. Imagine that the decisionmaker complies with certain axioms, which regiment the action and outcome rankings and how they intersect. Then these propositions will be true: (1) For each action a and each possible outcome x, there will exist a probability value $\pi_a(x)$, namely the probability that action a produces outcome x. (2) There will exist a utility function $v(.)$, which is unique up to a positive affine transformation, and which assigns utility values to outcomes $v(x)$. (3) The decisionmaker will have a complete ranking of the action set, which corresponds to the expected utility of the actions. EU theory, in this full form, will be discussed at length in Chapter 7.

EU theory did not originate with Savage. In earlier, seminal work, von Neumann and Morgenstern had introduced a simpler version of the theory, which applies to *lotteries* over outcomes. A lottery l over outcomes takes the form $[\pi_l(x), \pi_l(y), \ldots]$, where $\pi_l(x)$ denotes the probability that the lottery assigns to outcome x. Each such value $\pi_l(x)$ value is between zero and one, inclusive, and the sum of their values is one. A "lottery" is not an action, but an abstract, mathematical item, with probability values already "built in." The von Neumann/Morgenstern version of EU theory does not explicate what these probabilities are. In particular, it does not demonstrate that each particular action in an action set (if ranked by the decisionmaker in a rational manner) will correspond to a particular lottery across outcomes. Rather, what von Neumann/Morgenstern EU theory shows is this: If an individual is presented with a set of lotteries across outcomes, and ranks those lotteries consistent with a few simple axioms, then there will exist a utility function $v(x)$, unique up to a positive affine transformation, such that the individual's ranking of the lottery set will correspond to the expected utility of the outcomes. The von Neumann/Morgenstern axioms are *completeness* (the ranking of lotteries and outcomes must be a complete quasiordering); an *independence* axiom; and an *Archimedean* axiom.[83]

[83] Let **O** be a finite set of outcomes, and **L*** a set of all lotteries over the outcomes. (I denote this as **L*** to distinguish it from **L**, which is the set of all lotteries over *life-histories* that I discuss below.) The *completeness* axiom requires the decisionmaker to have a complete quasiordering of **L***. (Because an outcome is a "degenerate" lottery, assigning a probability 1 to the lottery, this means that the decisionmaker has a complete quasiordering of **O** as well.) The independence axiom involves the idea of "mixing" lotteries. A (p, 1-p) mixture of lottery r and lottery s assigns a given outcome a probability equaling p times the probability assigned it by r plus (1-p) the probability assigned it by s. What the independence axiom requires is that, if the decisionmaker weakly prefers l to l*, then she

Harsanyi employed the von-Neumann/Morgenstern version of EU theory, with a twist. Instead of thinking of outcomes as the possible "prizes" in each lottery, he thought of *life-histories* as the possible "prizes" in each lottery. Specifically, given some set **O** of outcomes, and some population of *N* individuals, we can define not just a set **H** of life-histories, but a set **L** of all lotteries over the life-histories. Each element of this set **L** is a lottery *l* which has the following form: [$\pi_l(x; 1)$, $\pi_l(x; 2)$, …, $\pi_l(x; N)$, $\pi_l(y; 1)$, $\pi_l(y; 2)$, …, $\pi_l(y; N)$, …]. The symbol "$\pi_l(x; i)$" represents the probability assigned by lottery *l* to life-history (*x*; *i*). A given lottery *l* will assign such a value to every life-history in **H**. The probability of any life-history must be a number between zero and one, inclusive. And the sum of the probabilities assigned by *l* to all the life-histories must be one. Set **L**, in turn, is the set of all possible such life-history lotteries.

Imagine, now, a given spectator *k*, who has extended preferences. Spectator *k* can rank the elements of the life-history set **H** and the elements of the life-history lottery set **L**. Further, assume that spectator *k* complies with the von Neumann/Morgenstern axioms: she has a *complete* ranking of **H** and **L**, and this ranking satisfies the "independence" and "Archimedean" axioms.

If so, there will exist a utility function $u^k(x; i)$, such that the spectator's extended preference ranking of the lotteries will correspond to the expected utility of the life-histories. In other words, the spectator will weakly prefer lottery *l* to lottery *l** iff the sum of the utility values assigned to each life-history by $u^k(.)$, discounted by its probability according to *l*, is at least as large as the sum of the utility values assigned to each life-history by $u^k(.)$, discounted by its probability according to *l**. Formally, there will exist a utility function $u^k(x; i)$, such that the

---

must weakly prefer a (*p*, 1-*p*) mixture of *l* with some lottery *r*, to a (*p*, 1-*p*) mixture of *l** with the same lottery *r* – for any *r* and *p*. Finally, what the Archimedean axioms says is that, if the decisionmaker prefers lottery *q* to *r* to *s*, then there is some mixture of *q* and *s* such that she is indifferent between this mixture and *r*.

If the decisionmaker ranks **L*** consistently with these three axioms, then there exists a utility function *v*(.), such that the decisionmaker weakly prefers *l* to *l** iff $\sum_{x\in\mathbf{O}} \pi_l(x)v(x) \geq \sum_{x\in\mathbf{O}} \pi_{l*}(x)v(x)$. This utility function is "unique up to a positive affine transformation," meaning that *w*(.) also expectationally represents the decisionmaker's preferences over **L*** (the decisionmaker weakly prefers *l* to *l** iff $\sum_{x\in\mathbf{O}} \pi_l(x)w(x) \geq \sum_{x\in\mathbf{O}} \pi_{l*}(x)w(x)$), just in case $w(x) = av(x) + b$ for all *x*, with *a* positive.

This version of EU theory assumes that the outcome set is finite. More generally, where the "prizes" are not outcomes (for example, where the "prizes" are life histories), this version of EU theory assumes a finite prize set. That will generally be my assumption throughout the chapter: **H** will be finite. EU theory also extends to the case of an infinite prize set, but this requires additional axioms, and may raise additional complexities in the case of extended preferences which I cannot address.

spectator weakly prefers l to l^* iff $\sum_{(x;i)\in\mathbf{H}} \pi_l(x;i)u^k(x;i) \geq \sum_{(x;i)\in\mathbf{H}} \pi_{l^*}(x;i)u^k(x;i)$. (The reader unaccustomed to this useful symbolism should consult the margin.) [84]

For short, if spectator k's extended preference ranking of **L** and **H** complies with the axioms of von Neumann/Morgenstern EU theory, there will exist a utility function which *expectationally represents k*'s ranking of the lottery set **L**.[85] Moreover, because a life-history is simply a "degenerate" lottery – a lottery that assigns probability 1 to that life-history, and zero to all others – this utility function $u^k(.)$ will also represent the spectator's extended preference ranking of **H**. It will be the case that the spectator weakly prefers $(x; i)$ to $(y; j)$ iff $u^k(x; i) \geq u^k(x; j)$. Finally, this utility function $u^k(.)$ will be unique up to a positive affine transformation. Some other utility function $v^k(.)$ will expectationally represent the spectator's preferences over **L** (as well as representing the spectator's preferences over **H**) iff $v(.) = ru(.) + s$, with r positive.

How does this construction relate to interpersonal comparisons? Harsanyi not only assumes that each spectator will have an extended preference ranking of life-histories and lotteries, consistent with EU theory, but adopts the further premise that spectators' extended preferences are *identical*. If so, there will be a single utility function $u(.)$, unique up to a positive affine transformation, which expectationally represents the preferences of all spectators. There will be no need to distinguish between the utility function $u^k(.)$ expectationally representing the preferences of spectator k, and a different utility function $u^j(.)$, which need not be a positive affine transformation of $u^k(.)$, and which expectationally represents the preferences of spectator j.

---

[84] What the symbol "$\pi_l(x;i)u^k(x;i)$" means is the probability assigned to life-history $(x; i)$ by lottery l, multiplied by the utility assigned by $u^k(.)$ to that life-history. What the symbol "$\sum_{(x;i)\in\mathbf{H}}$" means is that these values are calculated for each life-history belonging to **H** and summed. In other words, $\sum_{(x;i)\in\mathbf{H}} \pi_l(x;i)u^k(x;i) = \pi_l(x;1)u^k(x;1) + \pi_l(x;2)u^k(x;2) + \ldots + \pi_l(x;N)u^k(x;N) + \pi_l(y;1)u^k(y;1) + \ldots. \pi_l(y;N)u^k(y;N) + \pi_l(z;1)u^k(z;1) + \ldots + \pi_l(z;N)u^k(x;N) + \ldots$

[85] By "expectationally represents," I mean that $u^k(.)$ represents the spectator's preferences over lotteries in the manner just described, via the probabilistic expectation of $u^k(.)$: The spectator weakly prefers l to l^* iff $\sum_{(x;i)\in\mathbf{H}} u^k(x;i)\pi_l(x;i) \geq \sum_{(x;i)\in\mathbf{H}} u^k(x;i)\pi_{l^*}(x;i)$. It is quite possible for some $w^k(.)$ to represent the spectator's preferences over lotteries, but not *expectationally* represent them. For example, if $u^k(.) = h(w^k(.))$, with h not a positive affine transformation, then $w^k(.)$ will *not* expectationally represent the spectator's preferences over lotteries, but it *will* represent them via the rule: the spectator prefers l to l^* iff

$$\sum_{(x;i)\in\mathbf{H}} h(w^k(x;i))\pi_l(x;i) \geq \sum_{(x;i)\in\mathbf{H}} h(w^k(x;i))\pi_{l^*}(x;i)$$

If Harsanyi is indeed correct in his package of assumptions (that spectators possess extended preferences, that these are identical, and that these conform to von Neumann/Morgenstern EU theory), then the utility function $u(.)$ that arises in virtue of these assumptions is --- plausibly -- a kind of metric of the well-being *levels* of the different life histories. It is plausible to say that life-history $(x; i)$ is at least as good for well-being as life-history $(y; j)$ iff all the spectators weakly prefer the first to the second. Thus (granting Harsanyi his package of assumptions) it will be the case that life-history $(x; i)$ is at least as good for well-being as life-history $(y; j)$ iff $u(x; i) \geq u(y; j)$.

Of course, we don't really *need* the apparatus of EU theory to explicate comparisons of well-being levels. We could just say that life-history $(x; i)$ is at least as good for well-being as life-history $(y; j)$ iff all spectators weakly prefer the first to the second.

The real contribution of the EU apparatus has to do with well-being *differences* between the life-histories. If spectators rank the lotteries over the life-histories in compliance with the axioms of von-Neumann/Morgenstern EU theory, and have identical rankings that can be expectationally represented by a single utility function $u(.)$, then we can employ $u(.)$ to construct a *difference* ordering of the life-histories.

This possibility emerges because $u(.)$ is unique up to a positive affine transformation. Note that the family of utility functions consisting of $u(.)$ and all positive affine transformations of $u(.)$ generates a single, unique, difference ordering of the set of life-histories. Let us construct a utility function from $u(.)$ by saying that the difference between one pair of life-histories $[(x; i), (y; j)]$ is at least as great as the difference between a second pair $[(z; k) , (w; l)]$ iff $u(x; i) - u(y; j) \geq u(z; k) - u(w, l)$. This rule generates a complete difference ordering of the life-histories. Moreover – and this is the critical contribution of EU theory – the very same difference ordering of life-histories is generated by swapping $v(.)$ for $u(.)$, where $v(.)$ is a positive affine transformation of $u(.)$. Note that if $u(x; i) - u(y; j) \geq u(z; k) - u(w, l)$, it mathematically follows that $[ru(x; i)+s] - [ru(y; j)+s] \geq [ru(z; k)+s] - [ru(w, l)+s]$, with r positive.

In short, Harsanyi has sketched a possible route to constructing exactly what the SWF framework needs: a quasiordering of life-histories and of differences between life-histories, which allow for inter- as well as intrapersonal comparisons, and which are representable by utility numbers. For this contribution he deserves much credit.

However, as mentioned, I believe there are a number of critical gaps or flaws in Harsanyi's analysis. To begin, and most fundamentally, he does not confront the philosophical problem of explaining how an extended preference is possible. A preference is a choice-connected ranking. An individual's ordinary preferences are rankings that are linked to her choices, actual or hypothetical. What, then, does it mean for a spectator to have an extended preference – in particular, an extended preference regarding life-histories where she is not the subject?

Harsanyi suggests that a spectator should evaluate life-histories by "imagining himself to be *put in the place*" of the subjects. He suggests that spectator *k*'s extended preference ranking of (*x*; *i*) versus (*y*; *j*) involves comparing the prospect of being individual *i* in outcome *x*, to the prospect of being individual *j* in outcome *y*. But there are deep philosophical mysteries here. If the spectator and a given subject are different people, it is *impossible* for the spectator to *be* the subject. The spectator cannot, in any literal sense, "assume the identity" of a subject who is a different person from her. A very basic fact concerning the metaphysics of personal identity is that each person is necessarily identical to herself, and necessarily distinct from every other person. Individual *k* is the same person as individual *k*, and a different person from individual *i*, in each and every possible world.[86] Thus if spectator *k* is supposed to evaluate life-history (*x*; *i*) by imagining a state of affairs whereby individual *k* is identical to subject *i*, and outcome *x* occurs, then this ranking exercise requires spectator *k* to imagine something which is impossible. And the spectator's ranking of life-histories, thus construed, will be a "preference" only in the sense of being connected to "choices" that are impossible. Just as it is impossible for individual *k* to become individual *i*, so it is impossible for individual *k* to choose to become individual *i*. Thus, if spectator *k*'s preference for (*x*; *i*) over (*y; j*) is meant to be a "disposition" to chose an outcome in which *x* occurs and *k* becomes *i*, as opposed to an outcome in which *y* occurs and *k* becomes *j*, then this is a "disposition" to make one impossible choice over another.

The reader may protest that Harsanyi does not really intend that the spectator should imagine "assuming the identity" of the subject in any literal sense. Perhaps not. But then he owes us an explanation of what an extended preference ranking *does* involve, and how such a ranking is consistent with the metaphysics of personal identity. Neither Harsanyi, nor the subsequent economists or social choice theorists who have employed the idea of an extended preference, have done so.

Assume that we can make sense of extended preferences in a manner which is consistent with the metaphysics of personal identity. (I will argue that we can.) If so, we must use extended preferences to construct interpersonal comparisons in a manner that is sensitive to the "remoteness" objection to preference-based accounts of well-being. What if spectator *k* prefers life-history (*x*; *i*) to life-history (*y*; *j*) in virtue of facts about outcomes *x* and *y* that occur after the deaths of both subjects, or in some other way are too "remote" from these individuals to affect their well-being? Harsanyi does not confront this problem.

Moreover, Harsanyi's premise that spectators will have identical extended preferences is highly problematic. He presents an argument to back up this premise, but the argument is flawed, and has been powerfully criticized by John Broome and others. The critical premise in Harsanyi's argument is that there is a single set of causal laws that explains the variation in

[86] More precisely, individual *k* is self-identical in every world where he exists. Moreover, whatever we might say about the meaningfulness of the statement that *k* and *i* are not identical in worlds where one or both fails to exist, that statement is meaningful and true in every world where both individuals exist.

individual preferences. These laws, together with specific facts about any given individual, determine what preferences that individual will have: what her preferences regarding various types of actions and outcomes will be; and what her extended preferences will be. But even if this causal premise is true, and even if spectators are aware of the causal laws that predict individual preferences, the purported conclusion that spectator will have the same extended preferences over histories and lotteries is a non sequitur.

For example, imagine that outcomes are specified to include information about individuals' health and consumption. Imagine, further, that in outcome *x* one individual (Jim) has good health but little consumption, while in outcome *y* another individual (June) has poorer health but more consumption. Moreover, some spectators rank their own life-histories and life-histories involving other subjects so as to give substantial weight to consumption, while other spectators rank their own life-histories and life-histories involving other subjects so as to give more weight to health. In particular, some spectators have extended preferences that favor (*x*; Jim) over (*y*; June), while other spectators have extended preferences that favor (*y*; June) over (*x*; Jim). Add, now, the further fact that (1) each spectator knows what Jim and June's ordinary and extended preferences are; (2) each spectator knows what every other spectator's ordinary and extended preferences are; (3) each spectator knows the causal basis for (1) and (2). Does this further fact mean that, now, the spectators must converge in their ranking of (*y*; June) versus (*x*; Jim)? Why would it? The fact that the members of a community possess common knowledge regarding the diversity of their tastes, and the causal basis for the diversity of their tastes, is, clearly, not the same thing as the members having uniform tastes.

Further, even if spectators do have identical extended preferences over life-histories and lotteries, so that there exists a single $u(.)$ which expectationally represents these preferences, why assume that $u(.)$ is indeed the correct measure of well-being differences between the life-histories? Mightn't there be some other measure, which is a non-linear transformation of $u(.)$? One of the main themes in the critical literature regarding Harsanyi's scholarship is that he fails even to address this question. Although my position will be that something like $u(.)$ is indeed the correct measure of well-being differences between life-histories, I agree with the critics that the position needs a fuller defense than Harsanyi provides. He leaps too quickly from the sheer existence of a utility function that expectationally represents extended preferences over life histories, to the conclusion that this function measures well-being differences.

Finally, Harsanyi presents his ideas in the context of a defense of utilitarianism. As mentioned, Harsanyi assumes that each spectator develops a *moral* ranking of outcomes by construing each outcome as a equiprobability lottery over its component life-histories. This conception of what a moral ranking of outcomes involves leads to utilitarianism.[87] Harsanyi also presents a second, independent argument for utilitarianism, his so-called "aggregation theorem."

[87] See Chapter 4.

However, the concept of an extended preference regarding life-histories and lotteries is logically *separable* from utilitarianism. My account of well-being, below, will use that concept to build a set **U** of utility functions that represent the intra- and interpersonal ranking of life-histories and differences. It is a further and logically separate question, confronted in Chapter 4, whether **U** should be paired with a utilitarian SWF or some kind of non-utilitarian SWF. In Chapter 4, I argue for a *prioritarian* SWF.

This point bears repetition, since Harsanyi is famously associated with utilitarianism, and the reader may thus assume that anyone who uses Harsanyi's ideas must end up with that position as well. This is simply not the case. This book claims that (1) the well-being ranking of life-histories reduces to spectators' convergent extended preferences, but it *rejects* the further claim that (2) outcome x is morally better than outcome y just in case spectators converge in preferring an equiprobability lottery over outcome x's life-histories, to an equiprobability lottery over outcome y's.

Well-Being: An Account

This section draws upon the discussion of interpersonal comparisons in the previous section, as well as the discussion earlier in the chapter of philosophical debates about well-being, to present an account of well-being. The basic structure of the account is this: Given a life-history set and an associated set **L** of lotteries over the life-histories, each "spectator" k in the population of N individuals can be associated with a set $\mathbf{U}^k$: the set of utility functions that expectationally represent her fully informed, fully rational, extended preference ranking of **L** (more precisely, particular subsets of **L**) and that have nonexistence as the zero point. Pooling these individual sets across all spectators, we arrive at the set **U**.

In developing this account, I will of course need to provide an explanation of what an extended preference means – an explanation which is consistent with the metaphysics of personal identity. Again, this is a key issue which Harsanyi and other proponents of the idea of extended preferences have not confronted.

The account will be developed in a series of steps. First, I will begin the task of specifying a given spectator's extended preferences by specifying her "own-history" preferences: namely, extended preferences for life-histories in which she is the subject. Such preferences are metaphysically innocent. As we shall see, the idea of "self-interest" shall come into play in specifying a given spectator's "own-history" preferences. Next, I will discuss how the idea of own-history preferences can be *extended* to cover life-histories in which the spectator is not the subject. Third, I will discuss what it means for a spectator's extended preferences to be fully informed and fully rational. Fourth, I will discuss how the utility functions that expectationally represent a given spectator's preferences over **L** should be "zeroed out," so as to assign zero to nonexistence. Finally, I discuss the idea of pooling spectators' utility functions.

As we shall see, each spectator's extended preference ranking of **L** involves certain *counterfactual* choice situations: how she *would* rank certain hypothetical choices, if she were fully informed and fully rational. I do not believe that this feature of my account of well-being is a deficit. Questions about counterfactual choices can have determinate answers; and indeed the use of counterfactual choice situations to address questions about well-being or morality is widespread in philosophy. For example, full-information preferentialist views generally appeal to counterfactual rankings of outcomes and actions as the basis for well-being. The device of a hypothetical choice situation is a cornerstone of Rawls' approach to justice – principles for the basic structure of society are ascertained by asking which principles individuals would accept behind a hypothetical "veil of ignorance" -- and of the "social contract" tradition in moral thought that Rawls built upon.

Further, we should distinguish between the problem of *specifying* an account of well-being that invokes idealized extended preferences, and the problem of *estimating* well-being given that account. Tackling the first problem means identifying the particular hypothetical choice situations with reference to which the idea of an "extended preference" is to be understood, and making precise the concepts of "full information" and "rationality." A different question is how the nonomniscient moral deliberator (who will be using the account of well-being under construction here as a part of a moral decision procedure) should *estimate* *k*'s idealized extended preferences over **L**.

The problem of estimation is discussed at great length in Chapter 6. The focus, here, is the (logically prior) problem of specification.

(1) Own-History Preferences

We can begin the task of specifying a spectator's extended preferences by identifying his own-history preferences. A given spectator's "own-history preferences," as I will define them, involve life-histories in which he is the subject. For a given spectator *k*, consider the subset of the life-history set **H** containing life-histories of the form (*x*; *k*), (*y*; *k*), (*z*; *k*), …. And consider the subset of the set **L** of lotteries, such that any *l* within this subset assigns probability zero to a life-history in which *k* is not the subject. Then spectator *k*'s "own-history"' preferences are limited in scope to this subset of **H** and this subset of **L**.

But this specification of the scope of the spectator's own-history preferences does not, yet, provide a full definition of what they are. How do *k*'s own-history preferences relate to his ordinary preferences for outcome and choices?

It is tempting to say that *k*'s own-history preference regarding life-histories is just *identical* to his preference for outcomes: that *k* prefers (*x*; *k*) to (*y*; *k*) just in case *k* prefers outcome *x* to outcome *y*. But a spectator's preferences for outcomes may depend on features of outcomes that have nothing to do with his own well-being. If we define *k*'s own-history preferences as his preferences for outcomes, simpliciter, and then make such preferences part of

the basis for the well-being ranking of life-histories in which k is the subject, the resulting account of well-being will run afoul of the "remoteness" objection – even if the preferences are idealized to require full information and rationality.

I will therefore *define* a given spectator's own-history preferences as his *self-interested* preferences for the relevant outcomes. Spectator k has an own-history preference for $(x; k)$ over $(y; k)$ just in case he *self-interestedly* prefers outcome x to outcome y. Of course, we could achieve the same result by defining a spectator's own-history preferences as his preferences for outcomes, and then constructing an account of well-being that makes reference to spectators' "self-interested" own-history preferences. However, I think it is crisper to build the concept of self-interest into the very definition of an own-history preference.

What does it mean for spectator k to *self-interestedly* prefer outcome x to outcome y? As discussed earlier, I see two plausible approaches for defining "self-interest." One approach sees self-interested preferences as "value laden." It says: spectator k has a self-interested preference for outcome x over outcome y just in case k judges or perceives outcome x to be better for his well-being than outcome y. Another approach defines self-interest in terms of *sympathy* – understood as a type of psychological state which is not essentially value-laden. For spectator k to be sympathetic to individual j is for spectator k to have an attitude of care and concern towards j. Sympathy can be self-directed, or directed at others. And it can be partial or unreserved. On this approach to defining "self-interest," spectator k has a "self-interested" preference for outcome x over outcome y iff spectator, under conditions where he is unreservedly self-sympathetic, prefers outcome x to outcome y.

My account will be agnostic between these two approaches to defining "self-interest."

I have discussed own-history preferences for life-histories, but what about own-history preferences for lotteries? A lottery, as mentioned, is a mathematical item, assigning probability numbers to "prizes." More specifically, spectator k's own-history lottery l is a lottery of the form $[\pi_l(x; 1), \ldots \pi_l(x; k), \ldots, \pi_l(x; N), \pi_l(y; 1), \ldots \pi_l(y; k), \ldots \pi_l(y; N); \pi_l(z; 1), \ldots, \pi_l(z; k), \ldots, \pi_l(x; N), \ldots]$, with the special feature that $\pi_l(x; i)$ is given a non-zero value only if $i=k$. What the probability numbers in a lottery *mean* remains to be interpreted.

One standard interpretation sees a set of lotteries as a choice situation in which the decisionmaker's degrees-of-belief are fixed and exogenous. She enters the choice situation already possessing epistemic probabilities, measuring her degree of belief that any particular choice will yield a particular "prize." There has been some prior process (whatever it may be) yielding this assignment of epistemic probabilities to choices, and the decisionmaker now asks herself: Given that these are my degrees of belief, how do I rank the choices? On this interpretation, a particular lottery is just a numerical representation of a particular choice; and the probability numbers in the lottery are just the decisionmaker's exogenous degrees of belief that the choice will yield various prizes.

Building on this approach, I will define a given spectator's ranking of the subset of **L** containing his own-history lotteries as follows. Each lottery *l* corresponds to a choice in a hypothetical choice situation in which the spectator has exogenous epistemic probabilities. The probability values in a given lottery are the epistemic probabilities of the relevant outcomes. Specifically, own-history lottery *l* represents a hypothetical choice such that *k* believes to degree $\pi_l(x; k)$ that outcome *x* will occur; *k* believes to degree $\pi_l(y; k)$ that outcome *y* will occur; *k* believes to degree $\pi_l(z; k)$ that outcome *z* will occur; and so forth.[88] A different own-history lottery, *l**, represents a different hypothetical choice: a choice such that *k* believes to degree $\pi_{l*}(x; k)$ that outcome *x* will occur; *k* believes to degree $\pi_{l*}(y; k)$ that outcome *y* will occur; *k* believes to degree $\pi_{l*}(z; k)$ that outcome *z* will occur; and so forth.

Spectator *k* prefers lottery *l* to lottery *l** just in case he *self-interestedly* prefers the hypothetical choice corresponding to *l*, over the hypothetical choice corresponding to *l**. As before, I leave open whether "self-interest" should be defined in value-laden terms or in terms of a value-free attitude of unreserved self-sympathy.

(2) Extended preferences for life-histories in which the spectator is not the subject

I have specified a subset of a spectator's extended preferences, her own-history preferences, in a manner which uses the most familiar kinds of preferences, and which raises no difficulties concerning personal identity – by defining such preferences in terms of the spectator's preferences for outcomes and lotteries over outcomes, together with a self-interest component.

How shall we generalize this definition so as to make sense of *k*'s extended preferences regarding life-histories in which she is *not* the subject, and regarding lotteries in which the probability of life-histories in which she is not the subject is nonzero?

One possibility is to reduce a spectator's extended preferences to her own-history preferences, along the following lines. Imagine that life-history (*x; i*) and life-history (*z*; *k*) are related as follows: Subject *i*'s attributes in outcome *x* are identical to subject *k*'s attributes in outcome *z*. And imagine that life-history (*y*; *j*) and life-history (*w*; *k*) are related as follows. Subject *j*'s attributes in outcome *y* are identical to subject *k*'s attributes in outcome *w*. Then we might say that spectator *k* has a preference for life-history (*x*; *i*) over (*y*; *j*) just in case he has an own-history preference for (*z*; *k*) over (*w*; *k*).

This is a powerful idea. It will be the keystone of my *estimation* strategy, in Chapter 6: a strategy that uses existing data about an individual's preferences over outcomes and choices, to

[88] Because spectator *k*'s own-history lotteries assign zero probability to life-histories in which *k* is not the subject, an own-history lottery *can* be construed as a choice with fixed epistemic probabilities of outcomes. In other words, $\pi_l(x; i) = 0$ for any $i \neq k$ and any outcome *x*, and thus $\sum_{x \in \mathbf{O}} \pi_l(x; k) = 1$.

.

estimate her own-history preferences, and thereby to estimate her extended preferences more generally. Call this the "swapping attribute" approach to inferring a given spectator's extended preferences.

While the "swapping attribute" approach is a powerful inference tool, it is insufficiently general to serve as a *definition* of extended preferences.[89] The key problem has to do with subjects' *essential properties*. A person's essential properties are properties that he cannot lose, without losing his identity; they are properties that he possesses in every possible world where he exists. What individual properties are essential and what are contingent (non-essential) is a matter of debate, within the literature on personal identity. But, arguably, a person's essential properties might include her chromosomal makeup, the identity of her parents, or the date of her birth (if not the precise date, then the general time period). Arguably, it is impossible for Jim, who has chromosomes XY, and who was born in 2010, the child of Sam and Sheila, to have chromosomes XX, to be the child of other parents, or to be born in the year 500.

Imagine that we pursue the "swapping attributes" to defining extended preferences. Remember that a given outcome x has the form $(\mathbf{a}_1,\mathbf{a}_2,\ldots,\mathbf{a}_N,\mathbf{a}_{\text{imp}})$, where $\mathbf{a}_i$ are individual i's attributes, and $\mathbf{a}_{\text{imp}}$ are background facts. Imagine, now, that outcomes are specified so that each individual's attributes include not only her contingent attributes (her income, leisure, happiness, health), but also some of her essential attributes, whatever they may be. Imagine, further, that there are some subjects whose essential attributes are different from spectator k's. If so, the "swapping attribute" strategy collapses. That strategy says the following. If $x = (\mathbf{a}_1,\mathbf{a}_2,\ldots,\mathbf{a}_N,\mathbf{a}_{\text{imp}})$, then spectator k's extended preference for $(x; i)$ is equivalent to his own-history preference for some $(z; k)$, where $z = (\mathbf{a}_1^*,\mathbf{a}_2^*,\ldots,\mathbf{a}_N^*,\mathbf{a}_{\text{imp}}^*)$, and where $\mathbf{a}_k^* = \mathbf{a}_i$. But if individual i's essential attributes are different from spectator k's, there will not exist any such outcome z.

The proponent of the "swapping attribute" strategy might try to get around this difficulty by saying that spectator k's extended preference for $(x; i)$ is equivalent to his own-history preference for some $(z; k)$, where k's *contingent* attributes in z are identical to individual i's contingent attributes in x. But this "solution" would have a critical flaw: it would mean insisting, as a definitional matter, that a subject's essential properties cannot make any difference to his well-being.

I am not suggesting that subjects' essential properties do, in fact, have a substantial impact on their well-being. What I *am* suggesting is that this question should not be resolved as a conceptual or definitional matter. Seemingly, it is a conceptual *possibility* that individuals' essential properties do have some bearing on their well-being. And this seems to be a genuine possibility because of the room for debate about what, exactly, an individual's essential properties are. To pursue the "swapping attributes" approach to *defining* extended preferences

[89] Arguably, this was the definition of extended preferences that Harsanyi had in mind.

would rule out the very possibility that subjects' essential properties – whatever they may turn out to be – can have any impact on well-being. That seems unattractive.

I will therefore pursue a different approach – one that does not stipulate, at the outset, which kinds of properties (contingent versus essential) can influence a spectator's ranking of life-histories and lotteries in which he is not the subject.

Note, to begin, that we can readily generalize our definition of a spectator's own-history preferences to certain special subsets of **H** and **L**. Consider the subset of **H** consisting of all life-histories in which some individual *i* is the subject, where *i* is distinct from *k*. And consider the subset of **L** consisting of all lotteries in which the only histories with a non-zero probability have *i* as the subject. Let us call spectator *k*'s extended preferences for this particular subset of **H** and **L** his "*i*-history" preferences. Then we can define these preferences the very same way as we defined the spectator's own-history preferences, with one twist. Instead of being "self-interested," the spectator is now "*i-interested*": interested in the subject *i*. While the spectator's own-history preferences have been defined as his *self-interested* preferences for outcomes and hypothetical choices, the spectator's *i-history* preferences can be defined as his *i-interested* preferences for outcomes and hypothetical choices.

In other words: spectator *k* has an extended preference for life-history (*x*; *i*) over life-history (*y*; *i*) just in case spectator *k* prefers outcome *x* to outcome *y* when the spectator is *i*-interested. Spectator *k* has an extended preference for lottery *l* to lottery *l**, according non-zero probability only to *i*'s life-histories, just in case spectator *k* -- if he were *i*-interested --would prefer a hypothetical choice in which the epistemic probabilities of outcomes correspond to the probabilities in *l*, over a hypothetical choice in which the epistemic probabilities of outcomes correspond to the probabilities in *l**.[90]

What does it mean for a spectator *k* to be "*i*-interested"? We can answer this question using the very kinds of answers we used to define "self-interest." One possibility, the value-laden possibility, says that spectator *k* has an *i*-interested preference for outcome *x* over outcome *y* just in case *k* sees outcome *x* to be better for the well-being of *i* than outcome *y*. The other possibility invokes the value-free attitude of sympathy -- but now directed by the spectator onto subject *i*, rather than onto the spectator himself. On this construal of "*i*-interest," spectator *k* has an *i*-interested preference for outcome *x* over outcome *y* just in case the spectator, under a condition of unreserved sympathy towards subject *i*, prefers outcome *x* to *y*.

Just as I have been agnostic between the value-laden and sympathy based construals of self-interest, so I will be agnostic between the parallel construals of "*i*-interest." And just as my

[90] In other words, *k* sees *l*, a lottery in which only *i*'s histories have non-zero probability, as a hypothetical choice over outcomes such that the probability of outcome *x* is $\pi_l(x; i)$, the probability of outcome *y* is $\pi_l(y; i)$, and so forth.

definition of spectator *k*'s own-history preferences builds in a self-interest condition, so my definition of spectator *k*'s *i*-history preferences builds in a condition of *i*-interest.

I have thus succeeded in explaining what it means for spectator *k* to have an extended preference regarding a life-history in which he is not the subject, *without* requiring *k* to engage in metaphysically impossible thought experiments (imagining that he assumes the identity of someone else), and without making the implausible *conceptual* stipulation that such extended preferences are just equivalent to his preferences for own life-histories.

The strategy for specifying *k*'s extended preferences over the entire set **H** and entire set **L** will be to identify the rankings that *are* implied by his own-history and *i*-history rankings. We will thereby (if things go well) arrive at a single, complete extended preference ranking of **H** and **L**.

More precisely, imagine that a given spectator has own-history preferences, plus *i*-history preferences for each subject in the population other than himself. Further, imagine that these preferences are complete. Finally, imagine that these preferences comply with EU theory. Then there will exist a set $\mathbf{U}^k$, consisting of all utility functions that expectationally represent the spectator's own-history and *i*-history preferences. In other words, if $u^k(.)$ belongs to $\mathbf{U}^k$, then spectator *k* has a weak extended preference for own-history lottery *l* over own-history lottery *l** just in case the expected value of $u^k(.)$ with lottery *l* is at least as large as the expected value of $u^k(.)$ with lottery *l**. Spectator *k* has a weak own-history preference for life-history $(x; k)$ over $(y; k)$ just in case $u^k(x; k) \geq u^k(y; k)$. Spectator *k* has a weak extended preference for *i*-history lottery *l* over *i*-history lottery *l**, just in case the expected value of $u^k(.)$ with lottery *l* is at least as large as the expected value of $u^k(.)$ with lottery *l**. [91] Finally, spectator *k* has a weak extended preference for life-history $(x; i)$ over life-history $(y; i)$ just in case $u^k(x; i) \geq u^k(y; i)$.

If a spectator has complete own-history and *i*-history preferences, compliant with EU theory, there will exist such a set $\mathbf{U}^k$. How to construct it from the spectator's own-history and *i*-history preferences is discussed in the margin. [92]

---

[91] That is: (1) spectator *k* weakly prefers *l* to *l** iff $\sum_{(x;i)\in H} \pi_l(x;i)u^k(x;i) \geq \sum_{(x;i)\in H} \pi_{l*}(x;i)u^k(x;i)$, where *l* and *l** are own-history lotteries; and (2) the same is true where *l* and *l** are *i*-history lotteries.

[92] To create this set, consider the utility function $u_i^k(.)$, which is a utility function that expectationally represents spectator *k*'s preferences over *i*-history lotteries, assigning values only to life histories of the form $(x; i)$.
(The utility function $u_k^k(.)$ expectationally represents spectator *k*'s preferences over his own-history lotteries).
Denote as $\mathbf{U}_i^k$ the set consisting of $u_i^k(.)$ and all positive affine transformations thereof. Consider, now, the product set of these *N* sets. Each element of this product set is a vector of *N* utility functions, with the form $(u_1^k(.), u_2^k(.), \ldots, u_N^k(.))$. From each such vector of functions we can construct a grand function $u^k(.)$, by setting $u^k(x; i)$ equal to $u_i^k(x; i)$. Let $\mathbf{U}^k$ be the set of these.
It is clear that $\mathbf{U}^k$ does expectationally represent the spectator's own-history and *i*-history preferences, but it is *not* unique up to a positive affine transformation. The reason is this. If $u^k(.)$ belongs to $\mathbf{U}^k$ thus defined, consider $v^k(.)$, such that: (1) there is some particular subject *i*, such that $v^k(\_\_; i) = au^k(\_\_; i)+b$ for all life-histories, with *a*

Each $u^k(.)$ in $\mathbf{U}^k$, as constructed thus far, not only represents the subject's own-history and *i*-history preferences. It also implies a complete ranking of the entire set **H** and set **L**. It would be nice if every $u^k(.)$ in $\mathbf{U}^\mathbf{k}$ implied the very same complete ranking of the entire set **H** and entire set **L**. But this will *not* be the case.

It might be observed, here, that the grand set **U** which we are ultimately constructing in this chapter – the set which will represent the well-being ranking of life-histories and lotteries – can be constructed by pooling the various sets $\mathbf{U}^k$, without requiring that each $\mathbf{U}^k$ correspond to a single complete ranking of the entire set **H** and **L** by a given spectator *k*.

This observation is absolutely correct. However, if $\mathbf{U}^k$ is defined along the lines thus far discussed in this section (consisting of all utility functions that expectationally represent the spectator's own-history and *i*-history preferences), it may be so "large" that the set **U** created by pooling these sets across spectators may involve massive incomparability in the ranking of life-histories and differences. Can't we further "winnow down" $\mathbf{U}^k$?

Indeed we can. In order to do so, we need to introduce a new kind of extended preference: an extended preference on the part of spectator *k* for life-history (*x*; *i*) as opposed to life-history (*y*; *j*), where *i* and *j* are different persons; or an extended preference on the part of spectator *k* for a lottery *l* in which all the life-histories with non-zero probability are life-histories of subject *i*, as against a lottery *l** in which all the life-histories with non-zero probability are life-histories of subject *j* , where *i* and *j* are different persons.

Unfortunately, I see no way to define this new kind of extended preference in wholly value free terms. It is tempting to say that spectator *k* has an extended preference for (*x*; *i*) over (*y*: *j*) just in case the spectator, when unreservedly sympathetic towards the subject, prefers outcome *x* to outcome *j*. However, this seems meaningless when the subject is different in the two histories. It is meaningful to say that, when I am sympathetic towards myself, I prefer outcome *x* to outcome *y*. It is also meaningful to say that, when I am sympathetic towards Jim, I prefer outcome *x* to outcome *y*. But what does it mean to say that I prefer outcome *x* when I am sympathetic to Jim, to outcome *y* when I am sympathetic to Sheila? It is coherent to fix the spectator's valuational state (her information, rationality or, in this case, her attitudes of sympathy) and ask how she would rank outcomes with her mental states thus fixed. What seems incoherent is to ask how she would rank outcomes given a valuational state which varies along with the outcomes being ranked.

Instead, I will define this type of extended preference as follows: spectator *k* has an extended preference for (*x*; *i*) over (*y*; *j*), with *i* and *j* not identical, if spectator *k* judges that individual *i* has more well-being in outcome *x* than individual *j* does in outcome *y*. Similarly,

---

positive, and either *a* not equal to 1 or *b* different than zero; and (2) for all other subjects *j*, $v^k(\_\_; j) = u^k(\_\_; j)$. This new function is not a positive affine transformation of u^k but it *does* expectationally represent the spectator's own-history and *i*-history preferences and *does* belong to $\mathbf{U}^k$, as it has been constructed.

spectator *k* has an extended preference for a lottery *l* in which all the life-histories with a non-zero probability belong to subject *i*, over a lottery *l** in which all the life-histories with a non-zero probability belong to subject *j*, if *k* judges that the ex ante well-being for subject *i* produced by the hypothetical choice with outcome probabilities corresponding to *l*, is greater than the ex ante well-being for subject *j* produced by the hypothetical choice with outcome probabilities corresponding to *l**.[93]

It should be stressed that these "value-laden" definitions of the kinds of extended preferences now under analysis are consistent with pursuing a "value-free" approach to defining own-history and *i*-history preferences. Further, the relevant judgments of well-being need not be *thickly* value laden. In order to rank (*x*; *i*) versus (*y*; *j*), spectator *k* does not necessarily bring into play a wide range of value concepts. For example, he does not necessarily have in his head a whole list of objective goods. All that my definition requires is that spectator *k*, in some manner, reaches the minimally value-laden conclusion that subject *i* realizes more well-being in outcome *x* than subject *j* does in outcome *y*, or that subject *j*'s ex ante well-being with one choice is greater than subject *i*'s with another.

Assume that the spectator has complete own-history and *i*-history preferences. By adding information about his rankings of pairs of life-histories in which the subjects are different, or pairs of lotteries in which all the life-histories of the first lottery belong to one subject, and all the life-histories in the second belong to another, we can "winnow" down $\mathbf{U}^k$ – by excluding utility functions which are inconsistent with this new information. And (as elaborated in the margin) we will often, thereby, be able to "winnow down" $\mathbf{U}^k$ to the point where $\mathbf{U}^k$ does correspond to a single, complete, ranking of the entire set **H** and **L**. In this case, we will have "winnowed down" $\mathbf{U}^k$ so that its elements are unique up to a positive affine transformation. [94]

[93] By "ex ante" well-being, I just mean the well-being produced for someone by a choice whose outcomes may be uncertain. Note that, if the self-interest and *i*-interest conditions discussed earlier are construed in value-laden terms, the spectator must also make judgments of ex ante well-being in developing her own-history and *i*-history preferences. For example, the value-laden construal of a self-interested preference for own-history lottery *l* over own-history lottery *l** would be: the spectator judges that her own ex ante well-being produced by a lottery over outcomes with the probabilities specified by *l*, is greater than her own ex ante well-being produced by a lottery over outcomes with the probabilities specified by *l**.

[94] $\mathbf{U}^k$ can be thus "winnowed down" if there are at least two "points of contact" between the life-histories of individual *i* and individual *i*+1, for *i*=1 to *N*-1. (This is certainly *not* the only way to "winnow down" $\mathbf{U}^k$, but it is a sufficient condition for doing so.) To see how this works, consider a simple case of a population of 3 individuals (the idea generalizes to any finite population). Imagine that there are two outcomes, *x* and *z*, such that the spectator judges (*x*; 1) to be as good as (*z*; 2). Moreover, there are two other outcomes, *y* and *w*, such that: the spectator has an extended preference (a 1-history preference) for (*y*; 1) over (*x*; 1); she has an extended preference for (*w*; 2) over (*z*; 2); and she sees (*y*; 1) as equally good as (*w*; 2).
Choose some arbitrary $u^k(.)$ in $\mathbf{U}^k$. This function may already assign $u^k(x; 1) = u^k(z; 2)$ and $u^k(y; 1) = u^k(w; 2)$. If not, we can "rescale" it to do so as follows. Identify numbers *a* and *b*, *a* positive, such that $au^k(x; 1) + b = u^k(z; 2)$ and $au^k(y; 1) + b = u^k(w; 2)$. Then, specify a new function $v^k(.)$, component by component, as follows: $v^k(\_\_; 1) = au^k(\_\_; 1) + b$; $v^k(\_\_; 2) = u^k(\_\_; 2)$; $v^k(\_\_; 3) = u^k(\_\_; 3)$. Given how $\mathbf{U}^k$ has been constructed, $v^k(.)$ will be in $\mathbf{U}^k$.

Note that we will have thereby arrived at a single, complete, ranking of **H** and **L** without ever inquiring into spectator *k*'s preferences regarding lotteries that *mix* life-histories. A mixed history lottery *l* gives a non-zero probability to life-history (*x*; *i*), and a non-zero probability to life-history (*y*; *j*), where *i* and *j* are different individuals. It is not clear how to interpret a mixed history lottery. What kind of hypothetical choice, and what kind of value-free or value-laden attitude on the spectator's part, does a mixed-history lottery correspond to? My account does not suppose that a mixed-history lottery *does* correspond to a hypothetical choice. Rather, a mixed-history lottery, like all the elements of **L**, is a mathematical construct. Other elements of **L** do correspond to hypothetical choices. The spectator's ranking of these choices is represented by some $\mathbf{U}^k$, which in turn may well imply a ranking of all the elements of **L**, including mixed-history lotteries.

In some special cases, the spectator's complete own-history preferences and *i*-history preferences, together with the new information we have been discussing -- his ranking of pairs of life-histories involving different subjects and his ranking of pairs of non-mixed lotteries with different subjects – will not suffice to produce a $\mathbf{U}^k$ so that it represents a single, complete

---

Imagine, now, that there are also two "points of contact" between individuals 2 and 3. In other words, there are two outcomes, *e* and *g*, such that the spectator judges (*e*; 2) to be as good as (*g*; 3). Moreover, there are two other outcomes, *f* and *h*, such that: the spectator has an extended preference for (*f*; 2) over (*e*; 2); she has an extended preference for (*h*; 3) over (*g*; 3); and she sees (*f*; 2) as equally good as (*h*; 3). Identify numbers a^* and b^*, a^* positive, such that $a^*v^k(e; 2) + b^* = v^k(g; 3)$ and $a^*v^k(f; 2) + b^* = v^k(h; 3)$. Then specify a new function $w^k(.)$, component by component, as follows: $w^k(\_\_; 1) = a^*v^k(\_\_ ;1) + b^*$; $w^k(\_\_; 2) = a^*v^k(\_\_ ;2) + b^*$; $w^k(\_\_; 3) = v^k(\_\_ ;3)$.

This function $w^k(.)$ will also be in $\mathbf{U}^k$. Thus it will expectationally represent the spectator's 1-history preferences, 2-history preferences, and 3-history preferences. Moreover, it will represent the spectator's across-person judgments. That is, $w^k(x; 1) = w^k(z; 2)$ and $w^k(y; 1) = w^k(w; 2)$; $w^k(e; 2) = w^k(g; 3)$ and $w^k(f; 2) = w^k(h; 3)$.

Finally, it is straightforward to see that any other function in $\mathbf{U}^k$ which *also* represents these across-person judgments must be a positive affine transformation of $w^k(.)$. Why? Consider any $z^k(.)$ in $\mathbf{U}^k$. Given how this set has been constructed, there must be component-specific scaling factors that transform $w^k(.)$ into $z^k(.)$. In other words, there must exist $a_1, b_1, a_2, b_2, a_3, b_3$, such that: $z^k(\_\_; 1) = a_1w^k(\_\_; 1) + b_1$; $z^k(\_\_; 2) = a_2w^k(\_\_; 2) + b_2$; $z^k(\_\_; 3) = a_3w^k(\_\_; 3) + b_3$.

If $z^k(.)$ represents the spectator's across-person judgments, it must be the case that $z^k(x; 1) = z^k(z; 2)$ and $z^k(y; 1) = z^k(w; 2)$. Thus $a_1w^k(x; 1) + b_1 = a_2w^k(z; 2) + b_2$. And $a_1w^k(y; 1) + b_1 = a_2w^k(w; 2) + b_2$. Subtracting one equation from the other, we arrive at the conclusion that $a_1=a_2$, and thus $b_1=b_2$. Similar reasoning leads to the conclusion that $a_2=a_3$, and thus $b_2=b_3$. Thus $z^k(.) = a\, w^k(.) + b$.

Although I have showed how to "winnow down" $\mathbf{U}^k$ using across-person "points of contact" that compare life-histories, the idea straightforwardly generalizes to the case where these consist in lotteries over life-histories. In other words, there is some lottery *l* over individual 1's life-histories which is as good as some lottery *m* over individual 2's life-histories; and some *other* lottery l^* over individual 2's life-histories, which the spectator extendedly prefers to *l*, which is as good as some lottery m^* over individual 2's life-histories; and similarly for individuals 2 and 3, 3 and 4, and so on.

To be sure, in using the spectator's various across-person judgments to "winnow down" $\mathbf{U}^k$, these judgments have to be all consistent. "Consistency" can be most simply defined by saying just that there is at least one $u^k(.)$ in $\mathbf{U}^k$ that represents all these judgments. An inconsistent group of judgments will need to be revised.
.

ranking of **H** and **L**.[95] However, this possible eventuality is no insuperable obstacle to the account of well-being under development here. $\mathbf{U}^k$ will have been "windowed" down to some extent, at least. As in the case of a $\mathbf{U}^k$ that represents a complete ranking of **H** and **L**, it can be "zeroed out" (see below) and then pooled with other spectators' individual sets to create the grand set **U**.

(3) Full information and rationality

The philosophical literature is persuasive on the point that well-being has *critical* force, and thus that an account of well-being which is centered around preferences but fails to idealize those preferences in some manner is problematic. In particular, it is implausible that life-history (x; i) is better for well-being than life history (y; j) just because spectators, given their actual informational condition (perhaps quite imperfect) and state of rationality (perhaps quite poor), would have various self and other-regarding preferences over outcomes and choices, and these would imply utility functions that converge in giving a higher ranking to (x; i) than (y; j).

I therefore stipulate that the different spectator preferences discussed in the previous subsection, used to construct $\mathbf{U}^k$, should be "fully informed" and "fully rational." A spectator k has a fully-informed and rational own-history preference for life-history (x; k) over (y; k) just in case the spectator, under conditions of full information and rationality, would *self-interestedly* prefer outcome x to outcome y. A spectator k has a fully-informed and rational i-history preference for life-history (x; i) over (y; i) just in case the spectator, under conditions of full information and rationality, and under conditions where she is i-interested, would prefer outcome x to outcome y. Full information and rationality conditions can be similarly appended to the spectator's preference for lotteries over his own life-histories, for lotteries over someone else's life histories, and for comparisons of life histories or lotteries involving different subjects.

What do I mean by "fully informed"? Philosophers who have included a "full information" requirement in their favored account of well-being have specified this condition in a variety of different ways. There is a need to balance the critical force of additional information, with the fact that human beings do not have limitless cognitive abilities.

In my set up, the spectators are either ranking outcomes or hypothetical choices which correspond to lotteries – choices where the spectator's epistemic probabilities are fixed and exogenous. Thus we need not stipulate, and indeed should not stipulate, that the spectator knows for sure which outcome would result from which choice (for that would undercut the intendedly

[95] For example, imagine that there are 2 individuals and a set of outcomes, and the spectator judges that individual 1 is worse in each of his life histories than individual 2 is in any of his. Assume that $u^k(.)$ represents the spectator's own-history and i-history preferences, and is consistent with these across person judgments. Consider $v^k(.)$ such $v^k(\_; 1) = a_1 u^k(\_\_; 1) + b_1$, and $v^k(\_; 2) = a_2 u^k(\_\_; 2) + b_2$. Then it is quite possible that $a_1 \neq a_2$ and yet v^k is also consistent with all the across-person judgments. This is also possible where there is only 1 "point of contact" between the individuals' life histories, i.e., only one lottery (degenerate or non-degenerate) across individual 1's life-histories that the spectator judges to be equally good as one lottery (degenerate or non-degenerate) across individual 2's.

probabilistic nature of the choice-outcome nexus). Further, outcomes are simplified descriptions of realities, not complete possible worlds.

At one extreme, one might simply require that the spectators be fully informed regarding the stated characteristics of the outcomes. For example, if $x = (c_1, c_2, \ldots c_N)$, then k should be aware that individual 1's consumption level in x is indeed c_1, that individual 2's consumption level is c_2, and so forth. However, such information is probably too minimal. As Richard Brandt has shown through his detailed discussion of "cognitive psychotherapy," an attractive specification of a full-information condition should include information about the origin of an individual's preferences – for example, that he prefers outcome x because of social or parental pressure imposed on him as a child, or because of the association between his attributes in x and something pleasant.

I therefore tentatively propose the following specification of "full-information": each spectator is provided with (1) full information about the stated characteristics of the outcomes in **O**; (2) full information regarding the origins of his own preferences; and (3) any other information which he thinks relevant, if fully rational and either self-interested or *i*-interested, given (1) and (2). Under (3) might be information regarding the experiential quality of some feature of outcomes. For example, the description of some outcome, x, might describe individual i as having a particular health state in that outcome. Spectator k might never have experienced that health state himself, and might want more information about what the health state feels like – information which k might find relevant in formulating his i-history preferences.

Let us turn, then, to "full rationality." I distinguished earlier between procedural, historical and substantive specifications of "full rationality." At a minimum, "full rationality" on my account includes the following procedural requirements, very much in line with those suggested by other scholars who propose idealized-deliberation accounts of well-being or other normative constructs: the spectator's various rankings used to construct $\mathbf{U}^k$ must comply with the axioms of EU theory[96]; the spectator should not make mistakes of deductive or inductive reasoning; her first-order and higher-order preferences should be coherent; her attention should be focused on the ranking task at hand, rather than distracted by other issues; her emotional state should be calm.

Of course, as we've seen, various philosophers of well-being, have objected that an ideal-preference account which merely incorporates a procedural conception of full rationality is inadequate to screen out certain kinds of distorted preferences. For example, individual i might with full procedural rationality self-interestedly prefer outcome x, in which he has a low status and hardship-filled life, as opposed to outcome y, in which he has a more comfortable and higher-status life, because he has been indoctrinated to believe that he is a low status person and deserves nothing better .

[96] More specifically, she must comply with the von-Neumann/Morgenstern variant in choice situations where she has fixed, epistemic probabilities.

However, I am averse to specifying "full rationality" in partly non-procedural terms," for the following reasons. (1) On the account I am offering, the well-being ranking of life-histories depends on the preferences of all spectators. **U** is formed by pooling all the $\mathbf{U}^k$ sets. Individual *i*'s distorted preference for *x* over *y* will not suffice to make (*x*; *i*) better than (*y*; *i*), if other spectators (lacking the historical or other attributes that have distorted *i*'s preferences), have an *i*-interested preference for *y* over *x* and thus an extended preference for (*y*; *i*) over (*x*; *i*). (2) I have defined full information so that individuals are given full information about the causes of their preferences. In cases where the spectator's preferences are distorted in the historical sense, i.e., resulting from a process of indoctrination, undue social pressure, adaptation to terrible circumstances, and the like, the fully informed spectator will be made aware of the problematic origins of those preferences, and may end up with different preferences after fully procedural rational deliberation. (3) There has been little headway in specifying plausible historical conditions for preference rationality. (4) There has also been little headway in specifying plausible substantive conditions for preference rationality. There *are* various specific things which, intuitively, it seems to be substantively irrational to prefer. For example, if *k* prefers counting blades of grass, eating a saucer of mud, or feeling happy but only on Tuesday, this is (intuitively) irrational, even if *k* does so with full information, full procedural rationality, and an historically kosher process of preference formation. But little progress has been made providing a general rationale for the items that are substantively not preference-worthy.

(4) Zeroing out

Imagine that we have constructed a set $\mathbf{U}^k$ for some spectator *k*, along the lines discussed thus far. As discussed, we may often be able to "winnow down" the elements of $\mathbf{U}^\mathbf{k}$, so that it is unique up to a positive affine transformation. There is some utility function $u^k(.)$, such that $\mathbf{U}^k$ consists of $u^k(.)$ and all other utility functions $v^k(.)$ equaling $au^k(x; j) + b$, with *a* positive. If so, $\mathbf{U}^\mathbf{k}$ will correspond to a single, complete, ranking of the entire set **H** of life-histories and lotteries over life-histories. And it will also generate a unique ranking of differences between life-histories.

Note, however, that even if $\mathbf{U}^\mathbf{k}$ is unique up to a positive affine transformation, and a fortiori if it is not, $\mathbf{U}^\mathbf{k}$ will not imply a unique set of well-being ratios between life-histories. Imagine that $u^k(.)$ belongs to $\mathbf{U}^\mathbf{k}$, and that $u^k(.)$ implies that the well-being ratio between life-history (*x*; *i*) and life-history (*y*; *j*) is *r*. In other words, $u^k(x; i)/u^k(y; j) = r$. Consider now any utility function $v^k(.) = au^k(.) + b$, where *b* is any number other than zero. Then $v^k(.)$ is also a member of $\mathbf{U}^\mathbf{k}$. However, this utility function implies that the well-being ratio between the two life-histories, $v^k(x; i)/ v^k(y; j) = s = [au^k(x; i) + b]/[au^k(y; j) + b] \neq u^k(x; i)/u^k(y; j)$. For example, if $u^k(.)$ assigns the first life-history a utility of 10, and the second a utility of 5, then it implies that the first life-history has twice the well-being of the first. However, if $a = 3$ and $b = 15$, so that $v^k(.) = 3\,u^k(.) + 15$, this new utility function assigns the first life-history a utility of 45, and the second a utility of 30, implying a well-being ratio between two life-histories of 3/2.

Why be concerned about well-being ratios? There are various kinds of quantitative information about well-being that might be represented by utility functions: information about the well-being *levels* of life-histories, information about well-being *differences* between life-histories, and information about well-being *ratios* between life-histories. Earlier in the chapter, I provided a formal definition of what it means for information about well-being differences to be relevant to an SWF. Such information is relevant to many SWFs, but not to all (for example, difference information is not relevant to the leximin SWF).

Along similar lines, we can analyze whether information about well-being *ratios* is relevant to an SWF.[97] Not surprisingly, such information is not relevant to the leximin SWF. Nor is it relevant to the utilitarian SWF. But it is relevant to various other SWFs – in particular, to a continuous prioritarian SWF, as the following table illustrates.

[table]

Although this example involves a specific prioritarian SWF, using the square root function, it can be shown that ratio information is relevant to every continuous prioritarian SWF.

An intuitive argument for why ratio information is relevant to a continuous prioritarian SWF runs as follows. Any prioritarian SWF satisfies the Pigou-Dalton principle in terms of well-being. Decreasing a better-off person's well-being by a small amount, and improving a worse-off person's well-being by the very same amount, without affecting anyone else, is a moral improvement. But what if we arrange a "leaky" rather than perfect transfer of well-being between the two individuals? We decrease a better-off person's well-being by a small amount, and increase the worse-off person's well-being by only a fraction of that amount, for example one-tenth. Nine-tenths of the small amount "leaks away" in the transfer from the better to the worse off person.[98] Intuitively, our judgment whether this "leaky" transfer is morally attractive will depend on the well-being ratio between the two individuals. We will find it relevant, in reaching that judgment, whether the better-off individual is 20 times better off (in which case the leaky transfer may well seem morally attractive) , or only twice as well off (in which case it very well may not).

---

[97] By "relevant," here, I mean relevant in addition to information about well-being levels and differences. We can say, then, that ratio information is relevant to a given SWF if: (1) there exists some **U**, such that if we supplement **U** with $v(.)$, where $v(.)$ is a positive affine transformation of some member of **U**, the resulting set **U*** implies some range of ratios between at least one pair of life-histories that is not the same as that implied by **U**; and (2) the SWF, applied to **U***, yields a different quasiordering of outcomes than when applied to **U**. A positive affine transformation, of course, preserves level and difference information.

[98] The reader might worry that this example is circular. Doesn't the assumption that a certain *fraction* of the transfer "leaks" *already* imply ratio information? Aren't we just starting with ratio information to argue for ratio information? In fact, no. Imagine that **U** is unique up to a positive affine transformation. Then **U**, it turns out, uniquely defines *ratios of differences*. If $v(.) = au(.) + b$, a positive, then $[u(x; i) - u(y; j)]/[u(z; k) - u(w; l)] = [v(x; i) - v(y; j)]/[v(z; k) - v(w; l)]$. But it is not true that $u(x; i)/u(y; j) = v(x; i)/ v(y; j)$ unless b is zero. So we start with a **U** that gives complete information about well-being differences and levels and how "leaky" any transfer would be, and argue that it would be useful to "winnow" down this set so as to increase our information about well-being ratios.

How, then, shall we construct $\mathbf{U^k}$ so that it provides information about well-being ratios? Consider that the very idea of a well-being ratio involves a zero point. To say that the well-being of life-history (x; i) is five times the level of life-history (y; j) is just to say that the difference between the well-being of (x; i) and a zero point, divided by the difference between the well-being of (y; j) and a zero point, is five.

What, then, is the correct zero point for constructing well-being ratios? I suggest that the most attractive answer is *nonexistence*. There is no logical necessity in this answer. For a given life-history set, we could stipulate that some life-history (x^*; i^*) is the zero point, where the well-being of (x^*; i^*) might be better or worse than nonexistence; and we could "winnow down" $\mathbf{U^k}$ by including only functions that assign zero to (x^*; i^*). But it is very hard to see what rationale we could provide for identifying (x^*; i^*). By contrast, nonexistence quite naturally suggests itself as the zero point. Note, too, that there is a conceptual connection between the zero point and whether a life is "worth living." This is true, at least, if we represent whether a life is "worth living" in the most straightforward way, namely by assigning a positive number to a life worth living and a negative number to a life not worth living. If so, a life will be worth living iff it is above the zero point. But, presumably, a life is worth living iff it is better than *nonexistence*.

I therefore propose that we "winnow down" $\mathbf{U^k}$ by including only functions that assign zero to *nonexistence*. More specifically, imagine that there is some life-history where the spectator is the subject, (y^+; k), such that the spectator is indifferent between this life-history and nonexistence. In other words, the spectator under conditions of full information, full rationality, and self-interest, is indifferent between a state of affairs in which outcome $y+$ occurs and he (the spectator) exists, and his nonexistence. Some utility functions in $\mathbf{U^k}$ will assign zero to (y^+; k), others will not. The latter should now be excluded from $\mathbf{U^k}$.

Note that this procedure for "zeroing out" $\mathbf{U^k}$ requires the spectator to compare one of his own life-histories to nonexistence. So none of the subtleties involving a spectator's ranking of the life-histories of other subjects arise in this case. The spectator's own-history preferences, in general, are equivalent to his ranking of the corresponding outcomes under a condition of self-interest. The spectator's ranking of one of his own histories (x; k), vis a vis nonexistence, is equivalent to his self-interested preference ranking as between the outcome in the life-history (x) and nonexistence.

It might be protested that this impossible. It is impossible for the spectator to have a self-interested preference ranking of outcomes which includes not only various outcomes in which he exists (all the outcomes in $\mathbf{O}$), but also nonexistence. But why *is* this impossible? For a given spectator, think of nonexistence as some outcome n, additional to $\mathbf{O}$, where the spectator does not exist. It might be objected that the spectator cannot have a *preference* as between n and some outcome x in which he exists. A preference is not merely a ranking, but a ranking connected to *choices*; and there is no possible choice open to spectator k which would yield an outcome, n, in which k does not exist. Although this objection is cogent, the term "preference"

is sometime used in a broader sense to include favorable attitudes (what philosophers call "pro-attitudes") which are not necessarily choice-connected. In particular, a "wish" is a pro-attitude which need not be choice connected. Although the past is outside my causal control, I can have *wishes* regarding the past. I can *wish* that I didn't make that embarrassing comment to Suzy Q in 10th grade, even though there is nothing that I can do now to cause the nonoccurrence of the comment. I thus propose, more specifically, that the life-history (y^+; k), used to "zero out" $\mathbf{U}^k$, be such that the spectator, under conditions of full information, rationality, and self-interest, neither would wish the occurrence of outcome y^+ rather than outcome n, in which the spectator doesn't exist; nor wish the occurrence of outcome n rather than y.

For the remainder of the book, I will use the term "extended preference" to include the spectator's wishes regarding nonexistence. The other kinds of extended preferences used to construct $\mathbf{U}^k$ (the spectator's own-history preferences, *i*-history preferences, etc.) *are* genuine preferences. These are rankings which *are* connected to choices: choices that are hypothetical but still conceptually possible. The spectator's wishes regarding nonexistence are "preferences" only in the broader sense of being pro-attitudes, not in the stricter sense of being choice connected, but the reader will understand that in this particular instance (not generally) I am using "preference" in this broader sense.

A more technical difficulty is this: In a given outcome set **O**, there may not exist a particular life-history ($y+$, k), such that the spectator is indifferent between that history and nonexistence. This difficulty can be readily circumvented, by using the spectator's preferences (strictly, wishes) regarding lotteries that involve some probability of nonexistence. See the margin for details.[99]

What results when $\mathbf{U^k}$ is "zeroed" out, so that it includes only utility functions which assign zero to nonexistence? If $\mathbf{U^k}$, to begin, was unique up to a positive affine transformation, it is now unique up to a positive ratio transformation. $\mathbf{U^k}$ will consist of a single $u^k(.)$ and all positive multiples. If $v^k(.)$ belongs to $\mathbf{U^k}$, then it will be the case that $v^k(x; i) = c\, u^k(x; i)$, for all life histories.

If $\mathbf{U^k}$ is unique up to a positive ratio transformation, it will not only imply a single, complete, ranking of life-histories and a single, complete, ranking of differences between life-histories. It will also assign unique ratios between life-histories. Note that if $u^k(x; i)/u^k(y; j) = r$, and if $v^k(.) = cu^k(.)$, it follows that $v^k(x; i)/v^k(y; j) = r$. Moreover, even if $\mathbf{U^k}$, to begin, was *not* unique up to a positive affine transformation, the process of "zeroing out" may well mean that it

[99] Imagine that there is some life-history (x; k) and another (y; k), such that the spectator (under conditions of full information and rationality) prefers the first to the second, and prefers both to nonexistence. Determine the indifference probability p, such that k is indifferent between (y; k) and a lottery that gives him a p chance of (x; k) and a 1-p chance of nonexistence. A given $u^k(.)$ in $\mathbf{U}^k$ may be such that $u^k(y; k) = pu^k(x; k)$. In other words, $u^k(.)$ implicitly assigns zero to nonexistence. If it does so, it should be included, otherwise excluded from $\mathbf{U}^k$. A similar strategy can be used if there are two life-histories worse than nonexistence, or one better and one worse.

assigns unique ratios to some pairs of life-histories, or places the ratios between others within reasonably circumscribed ranges.

(5) The final step: Pooling life-histories

The prior sections specified a set $\mathbf{U}^k$ for each spectator. This set represents her fully informed, fully rational extended preferences regarding life histories, life-history lotteries, and comparisons to nonexistence.

The set **U** is formed by pooling these sets, across all *N* spectators. **U** is the union of $\mathbf{U}^1$ and $\mathbf{U}^2$ and … and $\mathbf{U}^N$. Just as each $\mathbf{U}^k$ is a theoretical construct – representing how individuals under idealized conditions would rank certain counterfactual choices – so, too, is **U** a theoretical construct. How to estimate the various $\mathbf{U}^k$ and, thus, **U** is discussed in Chapter 6.

If each $\mathbf{U}^k$ is unique up to a positive ratio transformation, *and* if spectators' fully informed, fully rational extended preferences are identical – if $\mathbf{U}^1 = \mathbf{U}^2 = \ldots = \mathbf{U}^N$ – then **U** will be unique up to a positive ratio transformation too. It will consist of some $u(.)$, and all positive multiples. In this simple, limiting case, **U** will define a *complete* well-being quasiordering of life-histories, and a complete difference quasiordering of life-histories. Remember that the rule for generating a well-being quasiordering of life-histories from a set **U** is: life-history $(x; i)$ is at least as good as life history $(y; j)$ iff, for all $u(.)$ belonging to **U**, $u(x; i) \geq u(y; j)$. And the rule for generating a difference quasiordering of life-histories from a set **U** is: the difference between life histories $(x; i)$ and $(y; j)$ is at least as large as the difference between $(z; k)$ and $(w; l)$ iff, for all $u(.)$ belonging to **U**, $u(x; i) - (y; j) \geq u(z; k) - u(w; l)$. If **U** is unique up to a positive ratio transformation, both of these quasiorderings will be complete. Moreover, **U** will ascribe a unique ratio to any pair of life-histories; and will yield a definite verdict regarding whether any life-history is better or worse than nonexistence.

Whether each spectator's $\mathbf{U}^k$ is unique up to a positive ratio transformation depends on some technical issues alluded to earlier. Whether spectators will have identical ("homogeneous") fully informed, fully rational extended preferences is a more basic question. I see no reason to assume, a priori, that they will. To begin, there is no reason to insist, a priori, that spectators' *non-ideal* extended preferences will be the same. [100] Do we secure homogeneity in extended preferences by idealizing them? Since "full rationality" has been specified in procedural, rather than substantive terms, there is no reason to assume that every fully informed, fully rational spectator *k* will end up with the very same set $\mathbf{U}^k$.

[100] For example, spectator *k* and spectator *l* might have different *i*-interested preferences over outcomes and choices, regarding some third person *i*. Spectator *k*'s *i*-interested preferences over outcomes and choices might differ from spectator *i*'s self-interested preferences. Finally, spectators *k* and *l* might differ in their across-person judgments (that one individual in some outcome is at least as well off as another individual in some outcome, or that one lottery across outcomes is better for one individual than another lottery is for another).

A great virtue of the account presented here, I believe, is that it *allows* for spectators to have identical extended preferences, but does not insist on such homogeneity. Homogeneity might be the case as an empirical matter; or it might be adopted as a simplifying premise, for purposes of estimation or otherwise. But where extended preferences *are* heterogeneous, the set **U** formed by pooling all the $\mathbf{U}^k$ will still be adequate as the basis for an SWF. It will still define an (incomplete) quasiordering of life histories and differences, allowing for some interpersonal comparisons of levels and differences (at least absent radical heterogeneity). And it will define ranges of ratios between the levels of different life-histories, if not unique ratios.

One distinction between the quasiordering of life-histories, and the quasiordering of differences, is worth noting. In the former case, we can see **U** as numerically representing a quasiordering that can be defined independent of **U**. We can say: life-history (x; i) is at least as good as (y; j) iff all spectators, under conditions of fully information and rationality, would weakly prefer (x; i) to (y; j). That, in turn, will be the case only if $u(x; i) \geq u(y; j)$ for all $u(.)$ belonging to **U**. By contrast, the difference quasiordering that I am arguing for cannot be defined independent of **U**. **U** is essentially involved in the very construction of that difference quasiordering. Of course, since **U** is essentially involved in the very construction of that difference quasiordering, **U** – trivially – represents the difference quasiordering

(6) Why is this an attractive account of well-being? Situating the account in the literature

I have proposed to analyze well-being in terms of fully informed, fully rational extended preferences. But why is this account attractive? Why should it be endorsed? My answer returns to the earlier sections of this chapter, surveying the philosophical literature on well-being and the case for interpersonal comparisons. I believe that the account fulfills a variety of desiderata outlined in these sections. Let me briefly summarize how it does so.

There is a powerful case for interpersonal comparisons. And the account, indeed, makes well-being interpersonally comparable. The extent of comparability depends on the degree of heterogeneity in spectator preferences. At the limit, where these are homogenous, and further each $\mathbf{U}^k$ is itself unique up to a ratio transformation, there will be full intra- *and* interpersonal comparability of life histories and differences. Every life-history will be better, worse, or equally good as every other life-history (regardless of whether the subjects are the same or different); and the difference between any pair of life-histories will be larger, smaller, or equal to the difference between every other pair.

The philosophical literature on well-being suggested, further, that an attractive account of well-being should have *critical* force; it should have *motivational* force; and it should not depend on features of outcomes too "*remote"* from the subject. My account is consistent with this trio of truisms. (1) *Critical force*. Because the preferences of a given individual k are *idealized* in the course of defining $\mathbf{U}^k$, and because the well-being ranking of k's life-histories depend on the extended preferences of the entire population, an individual can be wrong about her own well-

being. Individual k might incorrectly believe that $(y; k)$ is better for her than $(x; k)$ when, in fact, $u(x; k) > u(y; k)$ for all $u(.)$ in **U** and thus the second is better. (2) *Motivational force*. Well-being, thus defined, has motivating force under ideal conditions. If it is true that $(x; k)$ is better for individual k's well-being than $(y; k)$, then $u(x; k) \geq u(y; k)$ for all $u(.)$ in **U**, with at least one inequality strict. Given how **U** has been constructed, this implies, in turn, that $u^k(x; k) \geq u^k(y; k)$ for all $u^k(.)$ belonging to $\mathbf{U}^k$, i.e., that the subject k, under conditions of full information and rationality, would self-interestedly weakly prefer x to y. The well-being ranking of the subject's life histories is thus conceptually connected to her ideal *preferences*: to what she would be motivated to choose, if she had full information and rationality. Moreover, it is plausible that well-being has motivating force under some non-ideal conditions. Under some range of non-ideal conditions, individual k will be motivated by the knowledge that she would self-interestedly prefer x to y under ideal conditions. (3) *Remoteness*. Because $\mathbf{U}^k$ is built up from the spectator's self-interested and i-interested preferences for outcomes, it screens out features of outcomes that are too remote from the subject.

What kind of account is this? Is it a mental state account? A preference based account? An objective good account?

As I emphasized earlier, these are not mutually exclusive categories. My account makes reference, of course, to individual preferences. (Like other preferentialist accounts, such as the full information preferentialist views defended by Brandt and Railton, it does so in order to preserve the motivational truism.) But the account can also be seen as a kind of objective good account. A view of well-being is characterizable as an objective good account if it filters out idiosyncratic preferences -- if it does not reduce an individual's well-being to his own preferences, actual or idealized – and, further, does not make an individual's well-being a sole function of his mental states.

My account *is* an objective-good account in this sense. It analyzes the well-being ranking of life-histories, and differences between life-histories, in terms of individuals' *convergent* idealized preferences. One life-history is at least as good as a second iff *all* spectators, under conditions of full information and rationality, would weakly prefer the first to the second. This is true both when the subjects of the two histories are different, and when they are the same. For $(x; i)$ to be at least as good for well-being as $(y; i)$, it is a *necessary condition* that the subject under ideal conditions weakly prefer the first to the second; but this is not a sufficient condition.

Because the account appeals to *convergent* preferences, it has an appealing, "liberal," flavor. Philosophical liberals insist that individuals may have a diversity of conceptions of the good life, and that political theory and moral philosophy should respect such diversity. And indeed my account does respect such diversity. The well-being ranking of life-histories and differences represents the "zone of consensus" between individuals' rankings, which are permitted to be heterogeneous.

While the account is both a preference-based account and a kind of objective-good account, it cannot be characterized as a mental-state account. Nozick's experience machine provides a very strong case against insisting that well-being supervene on mental states. To be sure, the account I propose is consistent with the proposition that mental states are an important source of well-being, perhaps the dominant source. (If spectators' extended preferences depend largely on subjects' mental states, that will be the case). Indeed, the account is consistent with the proposition that mental states are, as a matter of contingent fact, the sole source of well-being. (If turns out to be the case that spectators' extended preferences solely depend on subjects' mental states, that will be true). But the account does not *necessitate* that mental states are the sole source of well-being. It is possible that (*x*; *i*) is better for well-being than (*y*; *j*) even though individual *i*'s mental states in *x* are identical to individual *j*'s mental states in *y*. The well-being ranking of these two histories depends on the various kinds of preferences used to construct each $\mathbf{U}^k$ and, ultimately, **U**. These preferences are restricted in various ways (they must be "self-interested," *i*-interested, and so forth), but these restrictions have been carefully defined so as to permit them to be sensitive to feature of outcomes other than the subjects' mental states. The extent to which spectators' extended preferences are thus sensitive is an empirical question.

Finally, how does the account relate to various metaethical positions? It sits most comfortably with an ideal-approval metaethics. In particular, it sits very comfortably with an ideal-approval view which analyzes moral facts in terms of individuals' convergent idealized preferences. Such a view says that outcome *x* is morally at least as good as outcome *y* iff all individuals, with full information and rationality, and taking an impartial perspective, would converge in weakly preferring *x* to *y*. It is natural to combine this view with the further claim that the *well-being* ranking of life-histories is a matter of individual's idealized preferences (now, not impartial preferences, but self-interested, *i*-interested, etc.) However, the account here is at least logically consistent with a much broader range of metaethical positions: for example, a noncognitivist view, or a cognitivist view which is distinct from an ideal-approval view.[101]

Lingering Objections

In the course of presenting the account, I have countered a number of potential objections; but there are several others, not yet discussed, which are worth addressing.

Intertemporal Change

In presenting the account, I assumed that each "spectator" is associated with a single fully informed, fully rational, extended-preference ranking of various subsets of **H** and **L**, giving rise to a single set $\mathbf{U}^k$. But it might be objected that these rankings can vary from time to time. Just as there can be intertemporal variation in an individual's ordinary preferences, so, too, we should allow for intertemporal variation in a spectator's extended preferences. Indeed, because a spectator's extended preferences are defined, in various ways, in terms of his ordinary

[101] **Discuss.**

preferences, it would be incoherent to allow for intertemporal change in ordinary preferences but not extended preferences. For example, spectator *k* has an extended preference for (*x*; *k*) over (*y*; *k*) iff he has a self-interested preference for outcome *x* over outcome *y*, and an extended preference for (*x*; *i*) over (*y*; *i*) iff he has an *i*-interested preference for outcome *x* over *y*. But spectator *k*'s preferences at one point in time regarding outcomes (be they self-interested preferences, *i*-interested preferences, or some other kind of outcome-preference) need not be the same as his preferences regarding outcomes at some other time.

This objection is quite cogent. Of course, as an *estimation* matter, it may well be simplest to assume intertemporal homogeneity in extended preferences; but the account itself should not insist on this. For simplicity, my presentation of the account in the prior section ignored the possibility of preference change; but the account can and should be generalized to allow for that possibility. Assume there are *T* time periods (most straightforwardly, the *T* periods in an outcome set specified in terms of multiple periods). Then spectator *k*'s fully informed and rational extended preferences at time *t* are captured in a set $\mathbf{U}^{k\text{-}t}$. With *N* spectators and *T* periods, there are *NT* such sets. Pooling all the sets, we arrive at **U**.

Sovereignty Respecting Preferences

An account of well-being might be described as "sovereignty respecting" insofar as it makes a given individual's own-history preferences, actual or idealized, determinative of the well-being ranking of those histories and differences between them. More precisely, an account might be termed "weakly" sovereignty respecting if it never *overrides* an individual's preferences regarding her own histories. And it might be termed "strongly" sovereignty respecting if it makes an individual's preferences regarding her own histories *decisive* in ranking the histories and differences between them.

The account I have proposed is weakly sovereignty respecting. If individual *i*, under conditions of full information and rationality, prefers life-history (*x*; *i*) to life-history (*y*; *i*), it will *not* be the case that the second life-history is ranked higher by **U**. This is because utility functions capturing individual *i*'s own-history preferences, via the set $\mathbf{U}^i$, are one component of **U**. Similarly, the account will never reach the conclusion that the difference between one pair of *i*'s life-histories is greater than the difference between another pair if $\mathbf{U}^i$ itself implies that the first difference is smaller.

However, this account is *not* strongly sovereignty respecting. If individual *i*, under conditions of full information and rationality, prefers life-history (*x*; *i*) to life-history (*y*; *i*), but some other individuals prefer (*y*; *i*) to (*x*; *i*), the upshot will be that **U** counts the two life-histories as incomparable. A similar result holds for differences. Indeed, my observation that the account is characterizable not only as a preference-based account, but also as a kind of objective good account, would not be true if it were strongly sovereignty respecting.

By contrast, Harsanyi employs extended preferences to arrive at an account of well-being that *is* strongly sovereignty respecting (as captured in his so-called principle of acceptance). This is one of the important differences between his account and my own.

My approach *could* be revised to make it strongly sovereignty respecting. This is clearly possible if each individual's preferences over her own histories and lotteries over those histories are intertemporally fixed and complete. Assume that everyone in the population is aware of everyone's own-history preferences. Then a given $\mathbf{U}^k$ could be constructed as follows. Spectator k should formulate her own-history preferences, as before, by looking to her self-interested preferences over outcomes and choices. But she should formulate her i-history preferences in a different way. Instead of defining those in terms of her i-interested preferences over outcomes and choices, spectator k's i-history preferences should just "piggy back" on individual i's own-history preferences. In other words, spectator k has an extended preference for $(x; i)$ over $(y; i)$ just in case individual i has an own-history preference for $(x; i)$ over $(y; i)$, i.e., just in case individual i would self-interestedly prefer outcome x to outcome y. Spectator k has an extended preference for one lottery over i's histories, as opposed to another lottery over i's histories, just in case individual i would have an own-history preference for the first lottery rather than the second.

If each $\mathbf{U}^k$ is constructed in this manner, and these sets are then pooled to create $\mathbf{U}$, the ranking of life-histories and differences with reference to $\mathbf{U}$ will be strongly sovereignty respecting. Complications arise if an individual's own-history preferences are allowed to vary over time, or to be incomplete. However, it seems plausible that these complications can be circumvented and a strongly sovereignty respecting $\mathbf{U}$ constructed even in these cases.

However, I have not pursued this approach, for several reasons. First, the philosophical case for making $\mathbf{U}$ strongly sovereignty respecting is hardly overwhelming. Objective good theories of well-being, which have substantial support in the literature, necessarily reject the proposition that well-being is strongly sovereignty respecting.

Second, a strongly sovereignty respecting $\mathbf{U}$ creates serious difficulties at the estimation stage – difficulties that are avoided by the account of well-being that I have proposed. At least at the margin, this sort of consideration is relevant in choosing between different accounts of well-being meant to serve as a component of a usable choice-evaluation framework. Such an account should be amenable to practical implementation, but a strongly sovereignty respect $\mathbf{U}$ is not. The difficulty, in a nutshell, is this. On the account that I have proposed, it is quite possible (although not mandated as a conceptual matter) that a given spectator's extended preferences depend solely on the subject's contingent attributes. In other words, each $u^k(.)$ in $\mathbf{U}^k$ may be such that $u^k(x; i)$ is determined by subject i's contingent attributes in outcome x. If so, we can infer $u^k(x; i)$ from spectator k's own-history preferences. Spectator k will prefer $(x; i)$ to $(y; j)$ just in case spectator k prefers $(w; k)$ to $(z; k)$, where w is such that k's attributes in this outcome are identical to i's in x, and z is such that k's attributes in this outcome are identical to j's in y. This is the heart of the

estimation strategy I employ in Chapter 6. However, if $\mathbf{U}^k$ is strongly sovereignty respecting, this simple and powerful approach to estimating individuals' extended preferences cannot be employed.[102]

Expected Utility Theory (Herein of incompleteness)

EU theory is an essential component of the account. The account assumes that each spectator's ranking of various hypothetical choices conforms to the simple von-Neumann/Morgenstern variant of EU theory.

Is this problematic? It is now quite widely accepted that EU theory fails to *describe* how individuals in many contexts actually behave; but I am employing EU theory here as a normative, not descriptive, theory. It is one aspect of the requirement that the spectator be "fully rational."

As we shall see in chapter 7, there *are* a variety of possible challenges to the normative credentials of EU theory. One important challenge says that EU theory lacks normative status for purposes of moral decisionmaking. Here, however, I am using the theory to structure spectators' *self-interested* or *i-interested* ranking of various choices, not spectators' moral ranking of choices.

A different, normative objection to EU theory says that, in some choice situations, individuals may have *indeterminate* probabilities. Indeed, this is true. But to say that indeterminate probabilities may occur in some choice situations is not, of course, to say that determinate probabilities can never occur. In particular, it is *possible* for a decisionmaker to face a choice situation in which she takes as given exogenous and determinate epistemic probabilities and asks herself: given that these are my degrees of belief, how do I rank the choices at hand? By specifying spectators' hypothetical choices in this manner, we can bring von Neumann/Morgenstern EU theory into play and construct a set $\mathbf{U}^k$ which will represent difference as well as level comparisons. It is hard to see how the sheer fact that this construction employs a hypothetical choice situation with exogenous and determinate probabilities is a problematic feature of the construction– whatever its other flaws.

A third objection involves *incompleteness*. EU theory assumes that individuals have complete rankings of outcomes and lotteries, but it is far from clear why this is normatively compelling. In particular, my construction of each $\mathbf{U}^k$ assumes that spectator k's self-interested preferences concerning outcomes and choices are complete, yielding a complete ranking of his

[102] Imagine that individuals i and j have different self-interested preferences regarding lotteries over attributes. Then if spectator k's $\mathbf{U}^k$ is constructed in a strongly sovereignty-respecting manner, it will need to distinguish between a life-history where i is the subject and the outcome is such that i has a certain attribute package, and a life-history where j is the subject and the outcome is such that j has the very same package. Otherwise, $\mathbf{U}^k$ will not be able to give a different ranking to pairs of attribute lotteries when these are lotteries over i's attributes, as opposed to when they are lotteries over j's attributes. In effect, $u^k(.)$ will need to assign utilities to life histories as a function of the subject's attributes *and* his name, not just his attributes.

own life-histories, and a complete ranking of all lotteries over his own life-histories. It also assumes that spectator *k*'s *i*-interested preferences concerning outcomes and choices are complete, yielding a complete ranking of *i*'s life-histories and complete ranking of all lotteries over *i*'s life-histories, for each subject *i*. Why believe that *k*, even if fully rational, would have complete rankings of this sort?

Indeed, a central thrust of this book has been to argue in favor of incomplete rankings of various kinds. I have claimed that the *moral* ranking of outcomes can be an incomplete quasiordering, and that the *well-being* ranking of life-histories and differences can also be an incomplete quasiordering. The assumption that each spectator's own-history and *i*-history preferences are complete helps to simplify the construction of **U**, and to facilitate the implementation of the account presented here, but this assumption would be problematic –in serious tension with my eagerness to allow other kinds of incompleteness-- if seen as a foundational feature of the account.

Fortunately, the assumption can be relaxed. Vitally important recent work in EU theory shows how it is possible to relax completeness but retain the other axioms of the theory. In particular, Efe Ok has demonstrated the following, in the von-Neumann/Morgenstern set-up: if an individual has an incomplete preference ranking of outcomes and lotteries, but her ranking satisfies the independence axiom and a continuity axiom, that incomplete ranking can be expectationally represented by a set of utility functions. The Ok result means that it is possible to construct a $\mathbf{U}^k$ which represents the spectator's extended preferences even if his own-history or *i*-history preferences are incomplete. The details are discussed in the margin.[103]

Should we use EU theory to construct a difference ordering?

An important theme in the critical literature regarding Harsanyi is that he provides no justification for using utilities that expectationally represent lotteries over life histories to measure the well-being differences between those histories. The objection, here, does not

[103] The Ok result shows that the spectator's incomplete preferences regarding lotteries over individual *i*'s histories (including the spectator's own), if consistent with the independence and continuity axioms, can be represented by $\mathbf{U}_i^k$. Each $u_i^k(.)$ in this set assigns a utility only to a life-history with subject *i*, one with the form (*x*; *i*). The spectator will weakly prefer lottery *l* over *i*'s histories to lottery *l** over *i*'s histories iff, for all $u(.)$ in $\mathbf{U}_i^k$,

$$\sum_{(x;i)\in\mathbf{H}} \pi_l(x;i)u_i^k(x;i) \geq \sum_{(x;i)\in\mathbf{H}} \pi_{l^*}(x;i)u_i^k(x;i).$$

Consider, now, the product set of these *N* sets. Each element of the product set is a vector of *N* functions, $(u_1^k(.), u_2^k(.), \ldots, u_N^k(.))$. For each such vector, define a new function $u^k(.)$ such that $u^k(\_\_\_; i) = u_i^k(\_\_\_; i)$. $\mathbf{U}^k$ is the set of all such functions. For any *i*, $\mathbf{U}^k$ thus constructed will represent the spectator's preferences regarding *i*-history lotteries, via the rule: *l* is at least as good as *l** iff, for all $u^k(.)$ in $\mathbf{U}^k$,

$\sum_{(x;i)\in\mathbf{H}} \pi_l(x;i)u^k(x;i) \geq \sum_{(x;i)\in\mathbf{H}} \pi_{l^*}(x;i)u^k(x;i)$. However, by contrast with the case where the spectator's preferences are complete, this $\mathbf{U}^k$ cannot be "winnowed down." The preferences are represented by the entire set, not by each member taken alone.

concern the general credentials of EU theory as an account of rational choice under uncertainty. Nor is the objection that well-being differences are not intra- or interpersonally comparable at all. (As we have seen, there is a strong case for difference comparability). The line of criticism I am referring to accepts the possibility of difference comparisons, and the generic credentials of EU theory, but challenges the use of EU theory to construct such comparisons.

Unlike certain other challenges to Harsanyi's views that have been frequently raised, this challenge can also be leveled at my account. The issue, here, is orthogonal to the question of homogenous versus heterogeneous extended preferences, and to the possibility of spectators having extended preferences that vary over time or are complete. I will therefore simplify the discussion by assuming that each spectator k has a set $\mathbf{U}^k$ which is unique up to a positive affine transformation and which, zeroed out, is unique up to a positive ratio transformation; and that these sets are identical across spectators. Thus **U** itself is unique up to a positive ratio transformation.

In general, of course, my approach *constructs* a difference ordering from **U**. In the case where **U** is unique up to a positive ratio transformation, this difference ordering will be complete. It will be the case that the well-being difference between life-history (x;i) and (y; j) is at least as large as the well-being difference between life-history (z; k) and life-history (w; l) iff $u(x; i)$-$u(y; j)\geq u(z; k)$-$u(w; l)$, where u(.) is any utility function in **U**. (Given uniqueness up to a positive ratio transformation, every u(.) in **U** will produce the same, complete, difference ordering). In particular, it will be the case that the well-being difference between life-history (x; i) and life-history (y; i) is at least as large as the well-being difference between life-history (z; i) and life-history (w; i) iff $u(x; i)$-$u(y; i)\geq u(z; i)$-$u(w; i)$.

At the same time, **U** will expectationally represent spectators' preferences over life-history lotteries. Consider any lottery l over some individual i's histories, and any other lottery l* over that individual's histories. Then, given how **U** has been "built up" from spectators' lottery preferences, it will be the case that spectators will prefer the first lottery to the second iff the expected utility from the first lottery, using u(.), is greater than the expected utility from the second lottery, using u(.). Of course, to say that u(.) expectationally represents lotteries over i's life-histories is, in turn, equivalent to saying that u(.) expectationally represents how spectators, concerned about i's well-being, would rank lotteries over outcomes.

But why think that there will necessarily exist a utility function which is able to play both these roles? [104] Why believe that we should be able to produce a u(.) which can do two things at once: (1) represent whether the difference between an individual's well-being in two outcomes is

[104] Strictly, my account doesn't say that there will necessarily exist such a u(.), just in the case where each spectator's $\mathbf{U}^k$ defines a complete ranking of the set of all life-history lotteries and the $\mathbf{U}^k$ are identical. But it is surprising that there should exist a u(.) which can play both of these roles even under these circumstances – even if spectators have the same complete extended preferences over lotteries.

larger or smaller than the difference between that individual's well-being in two other outcomes; and (2) expectationally represent how lotteries over outcomes should be ranked in light of that individual's well-being? Mightn't there be some measure $b(.)$, which is the true measure of well-being differences, and which need not equal $u(.)$ or a ratio or affine transformation thereof? If such a $b(.)$ existed, the relationship between $u(.)$ and $b(.)$ might look as follows.

[chart]

If the relation between $b(.)$ and $u(.)$ were non-linear in this way, spectators, ranking lotteries over outcomes in light of individual i's well-being, would be risk-averse or risk-prone in $b(.)$. The expected value of $b(.)$ would *not* correspond to the spectators' ranking of the lotteries.

By contrast, if $u(.)$ is itself the measure of well-being differences, as my account supposes – if $u(.)$ is just equal to $b(.)$ - the relation between $u(.)$ and $b(.)$ is necessarily linear. But to say that the relation between $u(.)$ and $b(.)$ is necessarily linear (as my account does) is to say that spectators, ranking lotteries over outcomes in light of individual i's well-being, must necessarily be risk-neutral in his well-being. Why believe that this is necessarily the case? It is a truism that the utilities provided by EU theory need not be linear in various sources of well-being. An individual can be "risk averse" or "risk prone" in terms of dollars, health, leisure, etc. Why shouldn't it also be possible for a self-interested individual, or a spectator concerned about that individual's well-being, to be risk averse or risk prone in well-being itself?

For this line of criticism of my account to be compelling, we would need to identify a competing method for arriving at a difference ordering of life-histories -- one that does not construct a difference ordering via spectators' preferences over life-history lotteries – and that is attractive on its own terms. However, I suggest that competing methods for producing a difference ordering are beset by various difficulties that my account avoids.

One possibility assumes that spectators can directly compare differences between life-histories. In particular, assume that each $\mathbf{U}^k$ is constructed as follows. Spectator k starts with his own life-histories and, as before, compares pairs of life-histories such as $(x; k)$ and $(y; k)$ by asking whether he self-interestedly prefers x or y. However, spectator k does not then proceed to think about lotteries at all. Instead, he thinks about whether the difference between $(x; k)$ and $(y; k)$ is greater than the difference between $(z; k)$ and $(w; k)$, for various such pairings. He does this by asking: Is the difference in my well-being as between x and y greater than the difference in my well-being between z and w?

Spectator k then engages in a parallel exercise with respect to the life histories of every other subject. Finally, k asks the sort of across-subject question anticipated by my account. He ranks pairs of life-histories such as $(x; i)$ and $(y; j)$ by asking: Is i's well-being in x greater than j's well-being in y? If things go well, all of these level and difference comparisons on the part of

k will be representable by a single utility function $u^k(.)$, unique up to a positive affine transformation.[105]

This approach is, necessarily, much more thickly "value laden" than the approach I have argued for. As discussed, my approach allows for the possibility that the spectator arrives at most (if not all) of his various rankings of life-histories and lotteries via value-free comparisons--comparisons that do not themselves involve judgments or perceptions of well-being. (For example, his ranking of lotteries over his own life-histories might be specified as a "self-interested" ranking of lotteries over outcomes, with "self-interested," in turn, understood, in value-free terms, as a condition of unreserved self-sympathy.) The approach I defend *also* allows for the possibility that the spectator arrives at his rankings of life-histories via value-laden comparisons. (The spectator's "self-interested" ranking of lotteries over outcomes *might* be understood as a judgment that one lottery is better for his well-being than a second.) By contrast, comparing pairs of outcomes with respect to differences in someone's well-being is necessarily a value-laden exercise. The spectator necessarily does this by thinking *about* the subject's well-being.

My account thereby preserves a kind of agnosticism with respect to a controversial question, rooted in unresolved metaethical disputes: to what extent do facts about well-being reduce to idealized value-laden judgments? It has the virtue of being reasonably robust to these disputes, while the competing approach now under consideration does not.

Moreover even if one is comfortable making spectators' rankings pervasively value-laden, it is a further question whether spectators can make judgments about differences. Such judgments (on the view under consideration) would have to be "primitive" judgments, which (1) are distinct from judgments concerning well-being levels, and (2) cannot be analyzed in terms of choices. As for (1): For spectator *k* to judge that individual *i* is at a higher level of well-being in *x* than *y*, and in *w* than *z*, leaves open whether the difference between *i*'s well-being in the first two outcomes is larger or smaller than the difference between *i*'s well-being in the second two outcomes. As for (2): If differences are not analyzable in terms of choices over lotteries, how does the spectator's judgment that the well-being difference between (*x*; *i*) and (*y*; *i*) is greater than the difference between (*z*; *i*) and (*w*; *i*) relate to any possible choice on her part? (Although one can choose an outcome, in the sense of selecting an action under a condition of certainty that the outcome would result from the action, it is not possible to choose an action yielding a difference between outcomes.). If the answer is negative, that direct difference comparisons have no connection to choices at all, once must then ask: are there really value judgments which are not choice connected?

[105] The literature on difference measurement shows that a complete difference ordering can be represented by a utility function unique up to a positive affine transformation. Thus spectator *k*'s ordering of the differences between individual *i*'s histories can be represented by all the members of some $\mathbf{U}_i^k$, unique up to a positive affine transformation. And we can generate $\mathbf{U}^k$ along the same lines as in note __ above.

Asking spectators to make direct difference comparisons is not the only competitor to the approach I have defended. There are a variety of other possible approaches, suggested by the literature. I cannot discuss these approaches at length. However, for reasons briefly reviewed in the margin, I suggest that these are even less attractive as a general methodology for constructing $\mathbf{U}^k$ than direct difference comparisons. (1) If a spectator's ranking of life-histories is assumed to be *temporally separable*, we might arrive at a $\mathbf{U}^k$ which is unique up to a ratio transformation, and which yields a complete difference ordering, without asking her to rank lotteries. (2) If a spectator's ranking of life-histories is assumed to be *separable across attributes*, we might again arrive at a $\mathbf{U}^k$ which is unique up to a ratio transformation, without asking the spectator to rank lotteries. (3) $\mathbf{U}^k$ might be built up from spectators' "just noticeable differences" in their ranking of life-histories. (4) The "zero-one" approach first proposed by Isbell might be employed to construct $b(.)$, namely by determining the utility function that expectationally represents each spectator's ordinary preferences regarding lotteries over her own life-histories, and then "scaling" the utilities thus produced so that 1 is assigned to each spectators' most preferred history and 0 to her least preferred.

To be sure, the reader might concede that the various competitors to the account proposed here are problematic, but still argue that some such competitor is, on balance, more attractive. "It is deeply counterintuitive that the correct measure of well-being differences is necessarily a linear function of $u(.)$. It is deeply counterintuitive that spectators, ranking lotteries across outcomes in light of someone's well-being, should be risk-neutral in that person's well-being." But is this really deeply counterintuitive? I don't believe that our intuitions on this score are that strong or well-formed. As Mattias Risse points out: "[O]ur intuitions about risk-attitudes with regard to well-being are not as developed as our intuitions about risk-attitudes with regard to, say, money …."The fact that an account of well-being entails risk-neutrality in well-being itself (as mine does) is an acceptable price to pay if the account is, on other dimensions, sufficiently attractive.